DIAGNOSTIC EFFICIENCY OF THE CHILD AND ADOLESCENT SYMPTOM INVENTORY (CASI-4R) DEPRESSION SUBSCALES FOR IDENTIFYING YOUTH MOOD DISORDERS

Stephanie Salcedo

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in The Department of Psychology and Neuroscience (Clinical Psychology).

Chapel Hill
2016

Approved by:

Eric A. Youngstrom

Deborah J. Jones

Andrea Hussong

ABSTRACT

Stephanie Salcedo: Diagnostic Efficiency of the Child and Adolescent Symptom Inventory (CASI-4R) Depression Subscales for identifying Youth Mood Disorders
(Under the direction of Eric A. Youngstrom)

This study examined the diagnostic and clinical utility of the Child and Adolescent Symptom Inventory (CASI-4R) depression and dysthymia subscales (caregiver and teacher report) for detecting youth mood disorders in outpatient mental health clinics (*N*=700). Semi-structured interviews (KSADS) with youth participants and their caregivers determined psychiatric diagnoses. CASI-4R depressive symptom severity and symptom count scores predicted mood disorder diagnoses. Both caregiver versions of the CASI-4R subscales significantly identified youth mood disorders (AUCs = .78 - .79, *p*s <.001). Teacher reports did not significantly predict mood disorder diagnosis (AUCs = .54, *p*s > .05). Caregiver subscale cutoff scores were calculated to assist in ruling in (DLR = 3.39) or ruling out (DLR = 0.36) presence of a mood disorder. The CASI-4R depression subscales caregiver report can help identify youth mood disorders. Using DLRs may help clinicians to identify youth mood disorders and improve diagnostic accuracy via these brief subscales.

ACKNOWLEDGEMENTS

I would like to thank Dr. Eric Youngstrom, Dr. Mary Fristad, the MECCA Lab (Mian-Li Ong, Tate Halverson, Yen-Ling Chen), and the LAMS team.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CASI-4R	Child and Adolescent Symptom Inventory – 4th Edition
K-SADS	Schedule for Affective Disorders and Schizophrenia for School-Age
DSM-IV-TR	Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision
RDoC	Research Domain Criteria
ROC	Receiver Operator Characteristics
AUC	Area under the Curve
DLR	Diagnostic Likelihood Ratio
CBCL	Child Behavior Checklist
TRF	Teacher Report Form
YSR	Youth Self Report
CDI	Child Depression Inventory
MFQ	Mood and Feelings Questionnaire
CAPA	Child and Adolescent Psychiatric Assessment
SDQ	Strength and Difficulties Questionnaire
LAMS	Longitudinal Assessment of Manic Symptoms study
PGBI-10M	Parent General Behavior Inventory-10-Item Mania Scale
ESM	Elevated Symptoms of Mania

INTRODUCTION

Depression affects a significant number of children and adolescents; epidemiological studies report three percent lifetime depressive disorder prevalence in preadolescence and up to 25 percent by the end of adolescence (Kessler, Avenevoli, & Ries Merikangas, 2001; Lewinsohn & Essau, 2002). When considering self-report, 20 to 50 percent of youths report depressive symptoms at clinically significant levels (Kessler et al., 2001). In childhood, there seem to be no differences in the rates of depressive disorders between boys and girls, but by early adolescence, girls' rates increase dramatically, to up to two times higher by age 15 (Hankin et al., 1998). This gender 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). Adolescents with depression are more likely to drop out of school or experience an unplanned pregnancy (Waslick, Kandel, & Kakouros, 2002). Furthermore, depression is the leading risk factor in youth suicide (Birmaher, Arbelaez, & Brent, 2002). Youths 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 (Lewinsohn, Rohde, Seeley, Klein, & Gotlib, 2003). Although most children and adolescents with a major depressive episode recover naturally, relapse rates are high, with the majority of youths experiencing another episode within several years later (Birmaher et al., 2002; Kovacs, 1996). Adolescents with major depressive disorder are at an elevated risk of having depressive episodes in adulthood (Birmaher et al.,

2002). 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. The goal of the present study is to evaluate a caregiver reported checklist for depressive symptoms and compare it to incumbent measures utilizing a sample that is generalizable to outpatient practice settings.

Assessment of Depression

Semi-structured diagnostic interviews. The "gold standard" to assess psychiatric disorders, including depression, is through a semi-structured diagnostic interview (Klein, Dougherty, & Olino, 2005). An interviewer, usually a trained researcher or clinician, will ask a client questions, make a rating based on the client's answers, and improvise any additional questions to clarify a client's answer or to address inconsistencies. Semi-structured interviews assess the criteria for most of the major child and adolescent psychiatric disorders, and when diagnosing a child, they also include a parent interview (Klein et al., 2005). Unstructured interviews are not preferred over structured interviews because clinicians may not consider all the necessary diagnostic criteria and overlook possible diagnoses; imposing structure reduces the likelihood of biases, such as selectively collecting information that confirms a clinician's initial impressions (Angold, 2002). In addition, semi-structured interviewing allows for the collection of more information beyond the standard interview questions, which may be relevant to understand the clinical presentation of a client. Incorporating clinical judgment with a structured interview is thought to increase the validity of the diagnostic interview (Kessler et al., 2001; Spitzer, 1983). The most widely used semi-structured interview for children and adolescents is the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS; Kaufman, Birmaher, Brent, Rao, et al., 1997). It has demonstrated fair to excellent interrater reliability for depressive disorders (Ambrosini, 2000; Kaufman, Birmaher, Brent, & Rao, 1997)

and has been used as the criterion for establishing validity with rating scales (Kaufman, Birmaher, Brent, Rao, et al., 1997).

Rating scales and checklists. Rating scales and checklists are another way to assess mood and symptoms of depression and serve as an initial screening tool. Depending on the scale, individuals rate the presence or the severity of a particular symptom in question, and summary scores from the symptoms can be calculated to gain an overall picture of the individual's symptom severity or impairment. Rating scales are attractive to clinical settings because, unlike structured diagnostic interviews, they are cost-efficient and do not require much time in terms of training or administration (Jensen & Haynes, 1986). Rating scales can have high face validity, especially because many of them are written to match with specific diagnostic criteria in mind (Jensen & Haynes, 1986). Furthermore, rating scales can be objectively scored and norms can be established to compare individuals' responses to others. Despite the benefits of rating scales, there are special considerations when assessing children and adolescents for depression. Youths might feel uncomfortable disclosing depressive symptoms or have trouble recalling the frequency, intensity, and duration of their specific symptoms (Garber & Kaminski, 2000). Therefore, incorporating rating scales that parents or teachers complete may be useful to collect information that youths do not report (Clarizio, 1994).

Incorporating multiple informants. When assessing for psychopathology in youths, collateral reports from various informants (e.g., parents, teachers) and settings (e.g., home, classroom) are often sought to gain a better understanding and more accurate picture of the youth's presenting symptoms. The added value for cross-informant reports varies depending on the age of the child; children are less reliable reporters of their psychological symptoms than adolescents (Edelbrock, Costello, Dulcan, Kalas, & Conover, 1985). Younger children may not

have the cognitive or linguistic capacity to process their experiences and report their subjective feelings, as well as recall temporal information such as age of onset or details of previous episodes (Edelbrock et al., 1985; Jensen et al., 1999; Kovacs, 1986). In addition, 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 (Richters & Pellegrini, 1989). Adolescents may also have issues reporting on the timeline and course of their symptoms, so parent and teacher report may still be useful (Edelbrock et al., 1985; Jensen et al., 1999).

However, it is important to consider that parents and teachers tend to report lower levels of depressive and other internalizing symptoms than youths may report about themselves (Jensen et al., 1999). Agreement between informants is between fair to moderate (Achenbach, McConaughy, & Howell, 1987). There are several potential reasons for this. First, parents and teachers may not observe the internalizing features of depression, which is why collateral reports are particularly useful with externalizing disorders (Klein et al., 2005). Second, parents with depression may have a lower threshold for noticing depressive symptoms in their children, so parent reports may result in a higher number of both true and false positive detections (Najman et al., 2000; Richters, 1992; Youngstrom, Izard, & Ackerman, 1999). Third, the inconsistency across informants may be a reflection of differences in how a child behaves in various settings, such as the home versus at school (Clarizio, 1994). However, there is still evidence for the validity of parent and child reports, and taken together, parent, teacher, youth, and clinician reports all explain unique and significant variance when predicting outcomes (Ferdinand et al., 2003; Verhulst, Dekker, & van der Ende, 1997). To integrate multiple informant reports, clinicians often use what is called the best-estimate approach, in which they integrate and weigh

the information from the various informants differently based on clinical judgment, using guidelines to increase the reliability and reduce bias (Klein et al., 2005). Other clinicians may use variations of the "or" or "and" rules, meaning that clinicians will take into account a symptom depending on whether the child presents this across multiple settings (Klein et al., 2005). An alternative is the "add" rule, in which clinicians, or often researchers, take the average scores of the self-report or collateral reports (Youngstrom, Findling, & Calabrese, 2003). Given that collateral reports are often used in a clinical setting with youths, we will compare caregiver and teacher reports on their performance in screening for youth depressive symptoms.

Gender differences in symptom presentation and assessment. Previous findings show a higher prevalence of depressive symptoms in girls relative to boys starting in adolescence (Hankin et al., 1998). Several theories have been established in the literature to explain potential reasons for this increase in depression among girls, such as the interactions between stress and biological changes associated with puberty (Cyranowski, Frank, Young, & Shear, 2000), girls' increased exposure to interpersonal challenges (Shih, Eberhart, Hammen, & Brennan, 2006), greater risk of exposure to traumatic sexual abuse (Hilt & Nolen-Hoeksema, 2009), as well as differences in how genders respond to stressful negative events, with girls tending to adapt a more internalizing and ruminative coping style (Nolen-Hoeksema, 2000).

Furthermore, another possibility is that boys and girls differ in their endorsement of their symptoms and/or how they respond to questions on measures. The few studies on measurement invariance and differential item functioning in children and adolescents have shown that there are differences in how boys and girls respond to items. In a study examining the Child Depression Inventory (CDI) in a sample of about 4000 school-age children and adolescents, van Beek, Hessen, Hutteman, Verhulp, and van Leuven (2012) found that girls were more likely to

endorse items examining sadness and crying as compared to boys. In addition, worrying about the future, self-blaming, and feeling like things bother them all the time were items endorsed at higher levels and more characteristic of depression in girls as compared to boys (van Beek et al., 2012). Similarly, Bares, Andrade, Delva, Grogan-Kaylor, and Kamata (2012) found that girls were more likely than boys to endorse depression items from the Youth Self Report (YSR). Therefore, research on measurement and assessment should closely examine whether scales perform differently in diagnostic prediction depending on gender.

The role of age in assessment. Rates of depression increase around age 14 in adolescence, particularly in girls (Hankin et al., 1998). Given that children's ability to process their experiences and understand their mental states continues to develop as they grow older, one question is whether younger and older children differ in how their depression is expressed or experienced. Previous research posits that although many core symptoms of depression seem to be the same across development, there are several symptoms that vary depending on age and developmental level (Cicchetti & Toth, 1998; Weiss & Garber, 2003). In the realm of assessment research, van Beek et al. (2012) found numerous instances of measurement invariance between adolescents and younger children; for instance, questions referring to self-esteem described the overall factor Self-Deprecation better for participants in adolescence as compared to those in elementary school. They also reported that at lower overall levels on the School Problems factor relative to elementary school children, early adolescent participants reported more difficulties with schoolwork (van Beek et al., 2012). However, less is known about how the type of informant influences the report of a youth's symptoms, and whether these age differences in symptom endorsement would be observed in caregiver reports. One possibility is that as the youth gets older, parent report of internalizing symptoms could diverge from the youth's

experience as a result of spending less time with each other as the child develops into adolescence.

Comorbidity as it complicates assessment of depression. Extensive evidence shows high rates of comorbidity of depression with other psychiatric disorders in youth. In a British community sample, 66% of children between the ages of five and 15 who had depression also had at least one comorbid diagnosis (Ford, Goodman, & Meltzer, 2003). In particular, there seems to be significant comorbidity between depression and anxiety disorders, conduct disorder, oppositional defiant Disorder, and ADHD (Angold, Costello, & Erkanli, 1999). Other studies have also found a link between depression and substance use disorders (Lewinsohn, Hops, Roberts, Seeley, & Andrews, 1993). Researchers have posited several reasons for the high rates of comorbidity with depression. Some have argued that it could be a reflection of the shortcomings of the current diagnostic system in that we are categorizing disorders into multiple subtypes that should belong in a single category, such as *internalizing disorders* or *affective disorders*. A second possibility is that one disorder may cause the development of another disorder (Cole, Peeke, Martin, Truglio, & Seroczynski, 1998). Rohde (2009) found that in 85% of youth cases with comorbid depression and anxiety disorders, the anxiety disorder seemed to have developed first. Furthermore, another argument is that the comorbid conditions are indicative of a distinct disorder (Harrington, Rutter, & Fombonne, 1996). Moreover, comorbid disorders could be a reflection of shared etiology (Klein, Riso, & Anderson, 1993).

One of the most prevalent comorbid diagnoses with depression are anxiety disorders, and many researchers have questioned whether these disorders are truly distinct (Watson et al., 1995). Clark and Watson's (1991) *tripartite model* conceptualized the relationship between anxiety and depression to account for the areas of overlap as well as their unique characteristics.

According to this model, anxiety and depression both contain negative affect, which is the extent to which an individual feels upset or emotionally distressed. Clark and Watson (1991) argued that self-report measures of anxiety and depression both measure negative affect, which can explain why they are highly correlated, and a scale may have difficulty discriminating between the two disorders if only negative affect is assessed. Where depression and anxiety differ is in levels of positive affect and physiological hyperarousal. Positive affect refers to overall positive or pleasant feelings about one's environment, and the absence of positive affect is conceptualized as anhedonia. Physiological hyperarousal is a state of increased physiological tension that can include somatic issues, shortness of breath, lightheadedness, among other problems. Elevated levels of anhedonia, or the absence of positive affect, distinctly characterizes depression, whereas elevated physiological hyperarousal is thought to uniquely distinguish anxiety from depression (Clark & Watson, 1991). More recent studies in children and adolescents have supported the tripartite model (Chorpita & Daleiden, 2002; Jacques & Mash, 2004; Lonigan, Phillips, & Hooe, 2003). When evaluating assessment tools for depression, it is important to consider the overlaps in depression and anxiety and focus on the development of screening tools that parse apart these disorders by focusing on their specific components.

Given the high rates of comorbidity of depressive symptoms with other disorders, one possible solution to address this heterogeneity is by using quantitative approaches to evaluate assessment tools transdiagnostically and categorize DSM diagnoses into groups that reflect common patterns or etiologies. This approach falls more in line with the Research Domain Criteria (RDoC) initiative proposed by the National Institute of Mental Health (NIMH), which proposes studying specific functions using multiple units of analysis and evaluate across diagnostic categories (Insel et al., 2010; Sanislow et al., 2010).

Calculating Diagnostic Efficiency

Screening tools can be helpful to determine whether someone is exhibiting symptoms consistent with a particular set of disorders; intuitively, one would expect that youths with depression would score higher on a depression screening tool than youths with no symptoms of depression. However, what is less clear is how a score on a particular scoring tool influences an individual's relative risk for a disorder. To translate research data into information that clinicians can utilize to answer this question, signal detection theory (McFall & Treat, 1999; Swets, Dawes, & Monahan, 2000) and Bayesian methods (Bayes & Price, 1763; Kruschke, 2011) provide the theoretical and statistical framework, specifically through the 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 the *sensitivity* and *specificity* of a test to evaluate diagnostic accuracy. Sensitivity is the proportion of individuals with a disorder who also received a positive test screening for that disorder, whereas sensitivity is the proportion of individuals without the disorder who received a negative test screening; the *false positive rate* is equal to 1 – 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 farther 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 utilize in their own practice, one could calculate diagnostic likelihood ratios (DLRs) from these scales' combination of sensitivity and specificity for a given score threshold. DLRs quantify the increased likelihood of having the disorder in question based on the individual's score on the scale. Furthermore, these approaches can be combined with probability nomograms, which 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 assessment and diagnosis, as well as streamlines successful treament by reducing clinician bias and incorporating multiple data sources (Jenkins, Youngstrom, Washburn, & Youngstrom, 2011; Youngstrom, Choukas-Bradley, Calhoun, & Jensen-Doss, 2015).

Prior Diagnostic Efficiency Research with Screening Tools for Depression

There are hundreds of measures for depression (Klein et al., 2005; Nezu, Ronan, Meadows, & McClure, 2000). However, a much smaller number of these measures have been examined using diagnostic efficiency statistics to determine the sensitivity and specificity in identifying depressive disorders. Table 1 shows available AUCs and DLRs for several of the depressive screening measures. We briefly review these measures and summarize what is known about their diagnostic accuracy.

Child Behavior Checklist and Youth Self Report. The Child Behavior Checklist (CBCL) is part of a multi-informant series of rating scales (Achenbach & Rescorla, 2001) that assess a broad range of psychopathology. In addition to the parent report, there is a Teacher

Report Form (TRF) and Youth Self Report (YSR); these measures are all comparable with one another (Hart & Lahey, 1999). The CBCL and TRF are now normed for youths ages 6 to 18, and the YSR is for youths ages 11 to 18. These scales have 118 items, and each item is rated on a three-point Likert scale, which takes 10 to 15 minutes to complete. The main factors these examine are anxious/depressed behavior, somatic complaints, withdrawn behavior, attention problems, social problems, thought problems, aggressive, and delinquent/rule breaking behavior. The reliability of the depression items is lower compared to other scales, and one argument is that these symptoms are less easily observable (Clarizio, 1994). Rather than use the anxious/depressed subscale, Lengua, Sadowski, Friedrich, and Fisher (2001) found that using an empirically-derived set of depression items from the CBCL showed greater sensitivity to major depressive disorder. The AUCs for the CBCL and YSR tend to be in the fair to good range (.70 to .81; Table 1).

Child Depression Inventory. The Child Depression Inventory (CDI; Kovacs, 1992) is a self-report questionnaire adapted from the Beck Depression Inventory (BDI; Beck & Steer, 1993) for children ages 7 to 17. The CDI contains 27 items, and individuals rate the severity of the listed depressive symptoms (with an emphasis on cognitive symptoms) over the past two weeks. The CDI takes about 10 to 20 minutes to complete, and a brief parent report version (Kovacs, 2003) is also available. Guidelines suggest the cut-off scores be 19 or 20 for youth in the general population or 12 to 13 for clinic-referred samples (Kovacs, 1992). The CDI has demonstrated high internal consistency (.80 in most studies; Crowley, Worchel, & Ash, 1992; Smucker, Craighead, Craighead, & Green, 1986). However, research on the test-retest reliability has been mixed, with some studies citing as high as.88 and others as low as .38 (Finch, Saylor, Edwards, & McIntosh, 1987; Saylor, Finch, Spirito, & Bennett, 1984). The evidence on the

CDI's discriminant validity has also varied. Timbremont, Braet, and Dressen (2004) found that the CDI can distinguish between depressed and nondepressed youth, but others suggest that it has poor discriminant validity because items from the CDI are highly correlated with measures of anxiety in addition to other depression measures, making it unclear the extent to which the CDI can distinguish youth depression from other disorders (Myers & Winters, 2002; Silverman et al., 1999). Sensitivity and specificity research for the CDI has revealed AUCs in the good to excellent range (.88 to .96; Table 1), suggesting the CDI is good at distinguishing youth depressive disorders from other disorders.

Mood and Feelings Questionnaire. The Mood and Feelings Questionnaire (MFQ; Costello & Angold, 1988) is a 32-item self-report questionnaire for children ages 8 to 18. Children rate the extent to which each statement is applicable to them on a scale of 1 to 3. The MFQ mainly covers DSM-IIIR (American Psychiatric Association, 1987) depression criteria and takes approximately 10 minutes to complete. The DSM-IIIR criteria for a major depressive episode required individuals to have "dysphoric mood or loss of interest or pleasure in almost all usual activities and pastimes" (Criterion A) followed by at least four Criterion B symptoms (similar to those in the DSM-IV-TR) for at least two weeks (American Psychiatric Association, 1987). The MFQ has good internal consistency, test-retest reliability, and strong convergent validity when compared to the CDI and diagnostic interviews such as the Child and Adolescent Psychiatric Assessment (CAPA) and K-SADS (Angold, Costello, Messer, & Pickles, 1995; Thapar & McGuffin, 1998; Wood, Kroll, Moore, & Harrington, 1995). There is also a parent version available, but some evidence suggests the diagnostic accuracy seems to be higher for the child version than the parent version when the K-SADS is used as the criterion (Wood et al., 1995).

Strengths and Difficulties Questionnaire. The Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) is a brief, 25-item measure of prosocial behavior and psychopathology in children and adolescents ages 3 to 16. Parent and teacher reports are available for 3-16 year olds, and self-reports are available for children 11 and over (Goodman, Meltzer, & Bailey, 1998). Respondents rate the extent to which each statement has been true for the youth in the past six months or last school year. The SDQ evaluates the following: emotional symptoms, conduct problems, hyperactivity/inattention, and peer relationship problems, and prosocial behavior. Computer algorithms predict the presence of a psychiatric disorder (conduct-oppositional disorders, hyperactivity, inattentive disorders, and anxiety-depressive disorders) by combining all available informant reports for the child (Goodman, Renfrew, & Mullick, 2000). The SDI is as good as the CBCL at detecting internalizing and externalizing problems (Goodman & Scott, 1999). The SDQ has also shown satisfactory reliability, as shown by internal consistency (.73), cross-informant correlation (mean: 0.34), and test-retest stability (.62; Goodman, 2001). AUCs for the SDQ in predicting depressive disorders are moderate (.78-.83; Table 1). However, these findings may be misleading and not generalizable to an outpatient clinic setting because two of these studies used community samples, so the prevalence and severity of depression is lower, and comparing a clinical diagnostic group with a group with no mental health issues is not likely to be a problem clinicians will be facing when evaluating a youth coming in to their clinic with mental health issues (Barrera & Garrison-Jones, 1988; Youngstrom, 2014).

The Current Research: Child and Adolescent Symptom Inventory

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, and its items correspond with DSM-IV criteria for the major childhood disorders. The CASI-4R combines the 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) subscale 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 utilizes 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.

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 regards to the depression and dysthymia subscales, preliminary research with the CSI-4 has found that the internal consistency is moderate for the parent version (.59 and .68) and high for the teacher version (.75 and .73; Gadow & Sprafkin, 2002). Convergent and discriminant validity have been examined comparing the CSI-4 parent and teacher reports to the CBCL and Teacher Report Forms, and the depression scales significantly correlated with the Withdrawn, Somatic Complaints, and Anxious/Depressed subscales of the CBCL and TRF, and the CASI dysthymia scales significantly 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 parent version of the CSI-4 in a sample of 247 clinically referred boys between the ages of six and 11; the results showed that the 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, they found that 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, more research is needed to determine how the CASI-4R (which includes both the CSI-4 and ASI-4) compares in its diagnostic performance, while also utilizing a sample of both males and females.

Current Study

Although several studies have examined the validity and reliability of the CASI-4R depressive and dysthymic disorder subscales for youth mood disorders, less is known about how an individual's scores on these subscales corresponds 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 13 seeking services in an outpatient mental health care setting. This study had several aims. **Aim 1** examined the diagnostic and clinical utility of the CASI-4R depressive subscales for detecting youth mood disorders. We hypothesized that both scoring methods (symptom count, symptom severity) will significantly discriminate between youth with current depressive symptoms from youths with other diagnoses. **Aim 2** compared the diagnostic efficiency of the symptom severity

and symptom count scoring methods. We expect that the symptom count scoring method will significantly outperform the symptom severity scoring method because the symptom count scoring method contains more items and covers a wider array of the criteria for depressive symptoms. We expected that the increase in content coverage and the larger number of items will offset the reduced variance per item, as the severity format scores items from 0 to 3, versus 0 or 1. **Aim 3** compared the diagnostic efficiency of the parent and the teacher reports of the CASI-4R. We hypothesized that the parent reports will significantly outperform the teacher reports, which is in line with previous research findings (Youngstrom et al., 2004). **Aim 4** examined gender and age differences in the diagnostic efficiency of the CASI-4R subscales. We expect that gender and age could moderate the effects of CASI scores on mood diagnosis, based on prior results, although the direction of moderation is ambiguous based on previous results. **Aim 5** demonstrated the clinical utility of the CASI-4R subscales through the use of a clinical vignette. Exploratory analyses also examined race as a covariate in predicting mood diagnosis.

METHODS

Participants and Procedures

These secondary analyses used baseline data from the Longitudinal Assessment of Manic Symptoms (LAMS) study (NIH R01MH073967-01; Coordinating PI: Robert L. Findling), 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 (Case Western Reserve University, Cincinnati Children's Medical Center, the Ohio State University, and University of Pittsburgh Medical Center/Western Psychiatric Institute and Clinic) approved all study procedures.

Parents/guardians of children between the ages of 6 and 12.9 years visiting one of the 10 child outpatient mental health clinics (two in Cincinnati, Ohio; two in Cleveland, Ohio; five in Columbus, Ohio; one in Pittsburgh, Pennsylvania) 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 who scored 11 or lower on the PGBI-10-M (ESM-), matched by age, sex, ethnicity, and Medicaid status, were also invited to participate to serve as the comparison group. To be screened for the study, participants must (1) not have previously 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. A detailed description of the screening procedures and rationale is described elsewhere (Horwitz et al., 2010).

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. After being assessed at baseline, participants came into the clinic every six months for five years, and each study visit lasted between two and four hours.

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, Birmaher, Brent, Rao, et al.,

1997), and they completed additional questions relating to depression and manic symptoms obtained from the Washington University in St. Louis Kiddie Schedule for Affective Disorders (WASH-U K-SADS; Geller, Warner, Williams, & Zimerman, 1998; Geller et al., 2001). Interviewer training consisted of attending a three-day start-up meeting, watching and rating taped interviews of the K-SADS-PL, and leading administrations of assessment instructions. Trained research assistants interviewed the youth participants and their caregivers, and a licensed clinical psychologist or psychiatrist reviewed K-SADS-PL data and diagnoses. The interrater reliability for the K-SADS-PL psychiatric diagnoses reached a kappa of 0.82 (Findling et al., 2010). Diagnoses were blind to CASI-4R scores.

Current depressive mood state was defined using thresholds from the Young Mania Rating Scale and Children's Depression Rating Scale (Duax, Youngstrom, Calabrese, & Findling, 2007). Using diagnoses derived from the K-SADS, we will create 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, bipolar I or II disorder (most recent episode depressed), or schizoaffective disorder (depressive type).

Broad definition. Children were included in this category if they 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 diagnosis category being analyzed. All participants were included in the analyses. All the participants were treatment-seeking individuals, so no healthy controls were included in the comparison group.

Child and Adolescent Symptom Inventory – 4R (CASI-4R; (Gadow & Sprafkin, 2005). The CASI-4R (163 items in parent report, 120 items in teacher report) examined the presence of DSM-IV criteria for emotional and behavioral disorders in children and adolescents. These analyses examined parent- and teacher-reported scores on the major depressive episode (MDE) and dysthymic disorder subscales. The MDD subscale contained 8 items in which symptoms related to DSM-IV criteria are described, and informants (parent and teacher) rated the frequency of the symptom on a scale of 0 to 3 (0 – *Never*, 1 – *Sometimes*, 2 – *Often*, 3 – *Very Often*). Sample items included, "[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 answered *yes* or *no* to an additional seven items related to the child's other depression related symptoms. These were not included in the standard severity score, but they contributed to the symptom count score. 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." The CASI-4R has shown satisfactory internal consistency, test-retest reliability, and theoretically consistent convergent and discriminant validity with respective scales from the Child Behavior Checklist (CBCL) and the Conners' Parent Rating Scale (Gadow & Sprafkin, 2005).

The CASI-4R MDE and dysthymic disorder subscales were scored in two ways, and we compared these two scoring methods to determine which scoring system worked best at predicting a depressed mood diagnosis.

Symptom severity scoring. For the first scoring option, the items scored 0-3 that closely matched with the DSM-IV criteria were summed and then averaged to get an average score for that subscale. The seven items included in this scoring method were: "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".

Symptom count scoring. For the second scoring option, the seven items previously described were recoded so that scores of 0 and 1 were recoded to 0 and scores of 2 or 3 were recoded to 1. These seven items will be summed and then averaged with the following *yes* or *no* items (where 0 – *no* and 1 – *yes*): "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)", "Has experienced a big change in his/her ability to concentrate or make decisions", "Feel that things never work out right", and "Has become more sensitive or tearful than usual".

Items that were not be included in either of the scoring methods were two items related to the criteria of causing significant distress or impairment, and one item assessing whether the child "has experienced a very stressful event such as parents' divorce, death of a friend or relative, serious illness." These three items were not included because they were not symptoms of depression per se, but rather associated features or tracking the presence of a precipitating event.

Data Analytic Plan

First, we examined baseline demographic and clinical characteristics, and we produced frequency distributions, measures of central tendency, variability, proportions, as well as boxplots and histograms to check for potential outliers. All tests used a two-tailed alpha-level of .05. Using independent samples *t*-tests, Levene's *F* tests, 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, broad, very broad) and children not in that category. Analyses were conducted using SPSS (Version 22.0) and the pROC and Aod packages in R.

Back-to-back histograms for all distributions detected any degenerate distributions or outliers. Next, nonparametric estimates of the area under the curve (AUC) from receiver operating characteristic (ROC) analyses (Youngstrom, 2014) quantified the diagnostic efficiency of the CASI-4R depression symptom severity and symptom count scoring forms (Aim 1). We calculated AUCs for the three diagnosis target categories (narrow, broad, and very broad) using both CASI-4R scoring methods to evaluate which scoring method is better able to discriminate youth with depressive symptoms from other youths in the sample. In addition, we calculated AUCs using both the caregiver and teacher versions of the scales. AUC guidelines suggest that

values less than .70 are considered poor, between .70 and .79 are fair, between .80 and .89 are good, and between .90 and 1.00 are excellent (Swets, 1988). However, it is important to note that values higher than .90 in mental health research may be indicative of a design flaw, such as using a comparison group that is too distinct (e.g., healthy controls; Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006).

Next, Venkatraman's permutation tests (Venkatraman, 2000; Venkatraman & Begg, 1996) compared ROC curves between the CASI-4R scoring forms, as well as the teacher and caregiver forms for each target diagnostic category (Aims 2 and 3). This test compared the 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 the AUCs appear the same at the second decimal point (Venkatraman, 2000; Venkatraman & Begg, 1996). For the CASI-4R scales with the highest diagnostic efficiency (caregiver or teacher version, symptom severity and symptom count scoring versions), hierarchical logistic regressions examined whether these scales still significantly predict the 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 scores as predictors of mood diagnosis. We also tested whether gender and age moderates the accuracy of the scale by calculating a gender by scale and age by scale interaction terms (Aim 4). Model 2 built off of model 1 by incorporating these interaction terms as predictors.

Next, we calculated diagnostic likelihood ratios (DLRs) from optimal cut-points to produce the best combination of the sensitivity and specificity from the ROC curves (Aim 5; Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, & Muller, 2011). The formula is as follows:

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

These DLRs provide clinically useful information regarding the 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). Lastly, we used a clinical vignette to show how to apply DLRs from this scale to an individual case using probability nomograms (Aim 5).

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 (Van Meter et al., under revision). Furthermore, weighting is frequently used in epidemiological studies for more accurate population parameter estimates.

RESULTS

Tables and results explained in this section refer to the broad target diagnosis category. Supplemental files (see Appendices A through D) display results for the narrow and very broad diagnosis categories.

Demographics and Preliminary Analyses

Table 2 presents the demographic and clinical differences between children with any broad mood diagnosis versus no broad mood diagnosis. Children in the broad mood group were significantly older ($p < .001$), more likely to be female ($p < .05$) and had more Axis I diagnoses ($p < .001$). Therefore, logistic regressions examined using these as covariates or potential moderators of CASI score accuracy. There were no group differences in race ($p = .12$). The broad mood group also had higher CASI caregiver and teacher scores (both scoring methods; *p*s < .004), which is in line with what we would expect because we grouped individuals by whether they met criteria for a broad mood diagnosis and are currently depressed. This is preliminary evidence that the CASI-4R depression and dysthymia subscales have discriminant validity.

Diagnostic Efficiency Analyses

Table 3 presents the AUCs from the ROC analyses for the CASI-4R caregiver and teacher reports of the 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 the caregiver reports were fair (Symptom severity AUC = .79, Symptom count AUC = .78, $p < .001$). Venkatraman's test compared the AUCs for the symptom severity versus symptom count scoring forms. The caregiver symptom severity scoring form outperformed the symptom count scoring form ($p < .001$). This suggests that the symptom severity scoring form of the caregiver report is better for detecting the presence of a depressive

disorder (using the broad diagnosis definition of depression). In contrast, AUCs for the teacher reports were poor and not statistically significant (Symptom severity and count AUCs = .54, *ps* > .05). Venkatraman's test revealed these teacher scoring forms were not significantly different (*p* > .05). See Figure 1 for ROC curves. When not including current depressive symptoms as part of the criterion definition for broad mood diagnosis, the results did not change substantively.

Covariate and Moderator Analyses

Logistic regressions examined whether the best-performing scale, the caregiver symptom severity scoring form, still significantly predicted the broad 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 scores were significant predictors of broad mood diagnosis after controlling for demographics and number of diagnoses ($B = 0.22$, $SE = 0.02$, $z = 9.72$, $p < .001$). This overall model explained 45% of the variance in broad mood diagnosis (Nagelkerke $R^2 = .45$). Age, number of diagnoses, and being female were significant predictors of broad mood diagnosis (*p*s < .0005). Being white did not predict broad mood diagnosis ($p = .10$).

Model 2 added sex by CASI and age by CASI interaction terms (Aim 4) to determine whether age or sex moderated the effects of CASI scores on broad mood diagnosis. The sex by CASI score interaction term was significant ($B = 0.12$, $SE = 0.06$, $z = 2.07$, $p = .05$), whereas the age by CASI score was not significant ($p > .10$). Venkatraman's test showed no significant difference in the AUCs between boys and girls (AUC = .82 versus AUC = .78, respectively, $p = .07$) This model explained 46% of the variance in broad mood diagnosis (Nagelkerke $R^2 = .46$). The main effects of age, number of diagnoses, and CASI scores were statistically significant in predicting broad mood diagnoses (*p*s < .01). The main effects of sex and race (being white),

however, were not significant (*p*s > .09). Model 2 explained an additional 1% of the variance in broad mood diagnosis ($p = .07$).

Diagnostic Likelihood Ratios (DLRs)

We estimated DLRs using cut-off scores that optimized sensitivity and specificity (Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, & Müller, 2011). Table 4 reports DLRs for the CASI-4R depression symptom severity scoring. Using optimal cut scores, a CASI depression severity score of 4.75 or higher resulted in moderate increase in odds (over three times) of meeting criteria for a broad mood diagnosis (*DLR* = 3.39). In contrast, scores below 4.75 were associated with reduced likelihood of receiving a broad mood diagnosis (*DLR* = 0.36). To more closely examine the sex by CASI score interaction, we compared optimal cut DLRs to see if they were different from one another. The DLRs for boys versus girls were within one point of each other, so although they were statistically significant, they were not clinically meaningful; having different DLRs translated to about 3-4% difference in the estimated probability, so for simplicity of application, we decided to keep the thresholds the same.

To provide more informative DLRs for clinicians to use, scores were divided into four thresholds (see Table 4). Youths who received a broad mood diagnosis were 7 times more likely to score 10 or higher on the CASI depression symptom severity subscale (*DLR* = 7.11). In contrast, individuals with CASI depression severity scores of 3 or below had decreased odds of meeting criteria for a broad mood diagnosis (*DLR* = 0.33).

DISCUSSION

The overarching goal of the present study was to examine the diagnostic efficiency and clinical utility of the CASI-4R depressive subscales for detecting youth mood disorders. We found that when using the caregiver report, both the symptom count and symptom severity

scoring methods significantly predicted broad mood diagnosis, but the symptom severity performed significantly better than the symptom count scoring method. This finding is inconsistent with our hypothesis that the symptom count scoring method would perform better possibly because of an increased coverage of symptoms with the larger item set. One possible reason why the symptom severity scoring method is resulting in greater diagnostic efficiency is we are provided with more information about each of the symptoms a child exhibits, rather than a simple indication of the presence or absence of a symptom. However, it is important to keep in mind that the difference in AUCs between these two scoring methods is .01, so it is unclear how clinically meaningful this difference truly is. However, one benefit to using the severity scoring method over the symptom count method is it eliminates the need to recode items, which leads to ease of use and application in a clinical setting.

We also found that both the scoring forms of the CASI teacher report did not perform above chance at predicting a broad mood diagnosis. 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). These findings make sense, given the CASI 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., change in sleeping patterns), so teachers in their limited time with a student may not be able to report changes in these areas.

The CASI severity scores still significantly predicted broad mood diagnosis after controlling for demographics and number of comorbidities. We found there was an interaction between CASI scores and sex, such that the CASI scores were marginally better at predicting broad mood diagnosis. However, comparing the likelihood ratios between boys and girls showed

that this difference was not clinically meaningful and may not warrant having separate ratios and cut scores by sex. The age by CASI scores interaction was also not significant at predicting broad mood diagnosis. Our study focused on diagnostic accuracy, so future research should examine whether differential item functioning is present in the CASI depending on the gender and age of the youth to determine whether children endorse symptoms differently depending on their characteristics.

The CASI depression AUCs for predicting a broad mood diagnosis are moderate. When comparing these results to AUCs for other scales, there are several other scales, such as the CDI, MFQ, and BDI, that seem to produce higher AUCs (see Table 1; for a review, see 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 the DLRs. Therefore, for clinicians to be able to come to a more accurate assessment, it is important to use DLRs based off similar settings (i.e., outpatient, inpatient, community mental health). One of the strengths of our study was that we conducted analyses from an outpatient clinical setting 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 compares in diagnostic efficiency in clinical settings to other established measures, such as the BDI, MFQ, and CDI.

One potential obstacle clinicians may face when choosing between administering the CASI and other scales is the cost. The CASI-4R is not free to use. However, the CASI-4R is not solely a depression measure; it contains subscales for all the DSM-IV childhood disorders. Therefore, although clinicians may have to purchase copies of the CASI-4R, they obtain access to scales for many disorders, unlike other measures that only examine one construct.

By including ROC analyses and DLRs in our analyses, it becomes easier for clinicians and to apply actuarial methods to 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 (for our last aim) 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 she has asked him why he has not completed his homework assignments, which has increasingly been more frequent. 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 easily distracted 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!" Celeste, however, initially thought that these behaviors were due to typical parent-child disagreements.

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 an 8.

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, you plot the DLR or Celeste's CASI score (from Table 3, you see it is 4.36) in the second line of the nomogram. You draw a line from the pretest probability value to the DLR and then extend the line to the posttest probability line. From there, you can see the posttest probability is about 43%. You can add a second nomogram to incorporate additional information, such as Jacob's family history of depression, by plotting 43% in the pretest probability line of the second nomogram. Then, you may decide to collect additional information, such as having Jacob complete a self-report scale or conduct a semi-structured interview with Jacob and Celeste. You can continue to integrate DLRs to adjust the posttest probability as shown above.

It is important to note that the 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, 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, then considering other potential disorders that may account for Jacob's behavior may be warranted.

Strengths and Limitations

This paper is the first to report AUCs and DLRs for the CASI-4R depressive subscales in both girls and boys, as well as in an outpatient sample. This study also used a sample of children with a ride range of diagnoses and no healthy controls, making the AUCs and DLRs more realistic to what clinicians may observe in their clinical setting. However, the 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 about a specific time point, 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 self-report scale was administered. Previous research suggests that caregivers are better reporter 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 is going through, their reports may not be as accurate as self-report as the child gets older and more aware of his/her mental state. Third, given the primary objective of the main outcome study was to track the progression of manic symptoms in children over time, ESM+ individuals comprise a large portion of the sample. To address this and make the findings more generalizable to other populations, we used weighting in our analyses, which is what is typically done in epidemiological studies. Further research about the use of the CASI in other settings, such as private offices or inpatient units is warranted and may improve our confidence in its use as a depression screener.

Conclusions

The caregiver CASI-4R depression subscales are useful tools at identifying youth mood disorders. The current results showed that elevated symptom severity scores are related to a 4 to

7-fold increase in the odds of meeting criteria for a broad 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 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, Calhoun, & Jensen-Doss, 2014).

Table 1

Areas Under the Curve (AUCs) and Likelihood Ratios for Screening Measures for Depressive Disorders

Screening Measure (Primary Reference)	**Citation**	**Reference Standard**	**AUC (N)**	**Optimal Cut-off Score**	**LR+**	**LR-**	**Population (age range)**	**Comparison group (Healthy controls: Yes/No)**
CBCL Anxious/Depressed Scale T-Score (Achenbach, 1991)	Nolan et al., 1996	Psychiatrist diagnosis	0.70	≥60	2.45[b]	0.64[b]	Subjects referred to six major health centers in Melbourne (ages 7-15)	Normative community sample from school-based asthma prevalence study
CBCL Anxious/Depressed Scale T-Score (Achenbach, 1991)	Eimecke et al., 2011	Multiaxal Classification of Child and Adolescent Psychiatric Disorders (MAS)	0.75	≥9	2.93[b]	0.52[b]	Adolescents referred for psychiatric services (ages 11-18)	No
CBCL Affective Problems Scale T-Score (Achenbach, 1991)	Eimecke et al., 2011	MAS	0.78	≥9	3.22[b]	0.45[b]	Adolescents referred for psychiatric services (ages 11-18)	No
CBCL Affective Problems Scale T-score (Achenbach, 1991)	Ferdinand, 2008	ADIS-P	0.77	--	3.13[b]	0.47[b]	Outpatient referrals for anxiety/depression problems (ages 6-18)	No

CDI Total Score (Kovacs, 1992)	Allgaier et al., 2012	German structured diagnostic interview for mental disorders in children and adolescents (Kinder-DIPS)	0.88	≥12	4.82[a]	0.20[a]	Inpatients and outpatients at pediatric hospitals ("medically ill children"; ages 9-12)	No
CDI Total Score (Kovacs, 1992)	Masip et al., 2010	Diagnosed by psychiatrists	0.93	≥19	23.75[a]	0.05[a]	Inpatients and outpatients at pediatric hospitals ("medically ill children"; ages 9-12)	No
CDI Total Score (Kovacs, 1992)	Timbremont, Braet, & Dreessen, 2004	KID-SCID	0.96[a]	≥16	14.96[a]	0.17[a]	Youths recruited from mental health center (ages 10-18)	Yes
CDI Short Form Total Score (Kovacs, 2003)	Allgaier et al., 2012	German structured diagnostic interview for mental disorders in children and adolescents (Kinder-DIPS)	0.88	≥3	3.17[a]	0.10[a]	Children and adolescents referred to inpatient or outpatient clinics (ages 8-18)	No
MFQ - Parent Report (Costello & Angold, 1988)	Daviss et al., 2006	K-SADS-PL	0.86	≥27	4.07[a]	0.46[a]	Recruited from bipolar offspring study and	Community controls

							psychiatric clinics (age 7 and up)	
MFQ- Child Report (Costello & Angold, 1988)	Daviss et al., 2006	K-SADS-PL	0.85	≥29	5.67[a]	0.36[a]	Recruited from bipolar offspring study and psychiatric clinics (age 7 and up)	Community controls
SDQ (Goodman, 1997)	Goodman et al., 2003	Developmental and Well-being Assessment (DAWBA)	0.83[b]	--	3.73[b]	0.32[b]	British Child Mental Health Survey (ages 5-15)	Community controls
SDQ (Goodman, 1997)	He et al., 2013	DSM-IV diagnostic information about their adolescent offsprings' mental health on self-administered questionnaire	0.78	≥13 (High Total Difficult.); ≥4 (Emotion Subscale)	4.26[a]	0.33[a]	National Comorbidity Survey – Adolescent Supplement (ages 13-18)	Community controls
SDQ (Parent; Goodman, 1997)	Johnson et al., 2014	DAWBA	0.79[a]	≥17	3.35[a]	0.41[a]	Extreme preterm children (ages 4-17)	No
YSR-CDI (combo of these two measures; Achenbach, 1991)	Rey et al., 1992	Two independent clinicians made separate DSM-III diagnoses	0.81	--	3.52[b]	0.37[b]	Referred adolescents to assessment in psychiatric hospital (ages 12-16)	No

YSR (Achenbach, 1991)	Ferdinand, 2008	ADIS-P	0.77	--	3.13[b]	0.47[b]	Outpatient referrals for anxiety/depression problems (ages 6-18)	No

Note:

"LR+" refers to the change in diagnostic likelihood ratio associated with a positive test score (above the score cut-off), and "LR-" is the likelihood ratio for a score below the score cut-off threshold.

[a]Values calculated based on sensitivity and specificity values provided in study

[b]Estimated values (Calculated from AUCs and sensitivity values at a given a specificity level of 0.8 using formula from Hasselbad & Hedges, 1995).

Table 2

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

Variable		Any Broad Mood (*n* = 259)	No Broad Mood (*n* = 441)	Test Statistic	*p*	Effect Size
Age	*M*	9.91	9.05	*t*(698)=5.91	<.001	*d* = 0.45
	SD	1.91	1.83	Levene's *F* = 1.70	.19	
Female		*n* = 99 (38%)	*n* =127 (29%)	X^2 (1) = 6.63	.01	phi = 0.10
Race (White %)		*n* = 176 (68%)	*n* = 274 (62%)	X^2 (1) = 2.41	.12	phi = 0.05
Axis I Diagnoses at Baseline	*M*	3.17	2.10	*t*(464.4)=11.11	<.001	*d* = 0.84
	SD	1.31	1.08	Levene's *F* = 12.18	.001	
CASI-4R Symptom Severity - Caregiver	*M*	7.31	3.63	*t*(428.0) = 10.98	<.001	*d* = 0.85
	SD	4.55	3.60	Levene's *F* =15.19	<.001	
CASI-4R Symptom Severity - Teacher	*M*	4.59	3.37	*t*(460) = 3.18	.002	*d* = 0.30
	SD	4.32	3.79	Levene's *F* = 2.64	.11	
CASI-4R Symptom Count - Caregiver	*M*	3.82	1.95	*t*(415.3)=9.12	<.001	*d* = 0.73
	SD	2.82	2.14	Levene's *F* =31.70	<.001	
CASI-4R Symptom Count - Teacher	*M*	4.40	3.34	*t*(318.8)=2.92	.004	*d* = 0.29
	SD	3.91	3.43	Levene's *F* =4.10	.04	

Note: Where data points were missing, effect sizes were calculated out of total number of available cases.

Table 3
Areas Under the Curve from Receiver Operating Characteristic analyses for Broad Mood Disorder Status

		95% Confidence Interval		
Index Test	**Area Under Curve**	**Lower**	**Upper**	**Difference between AUCs**
CASI-4R Symptom Severity - Caregiver	.79***	.76	.82	$p < .001$
CASI-4R Symptom Count - Caregiver	.78***	.75	.81	
CASI-4R Symptom Severity – Teacher	.54	.50	.58	$p > .05$
CASI-4R Symptom Count – Teacher	.54	.50	.58	

Note: *$p<.05$, **$p<.005$, ***$p<.001$, two tailed.

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

Cut Score	Level	DLR
Multilevel DLRs (based on optimal thresholds)		
0 to 4.74	64%	0.36
4.75 and above	36%	3.39
Multilevel DLRs (based on more informative thresholds)		
0 to 3	59%	0.33
4 to 6	18%	1.27
7 to 9	12%	4.36
10+	11%	7.11

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.

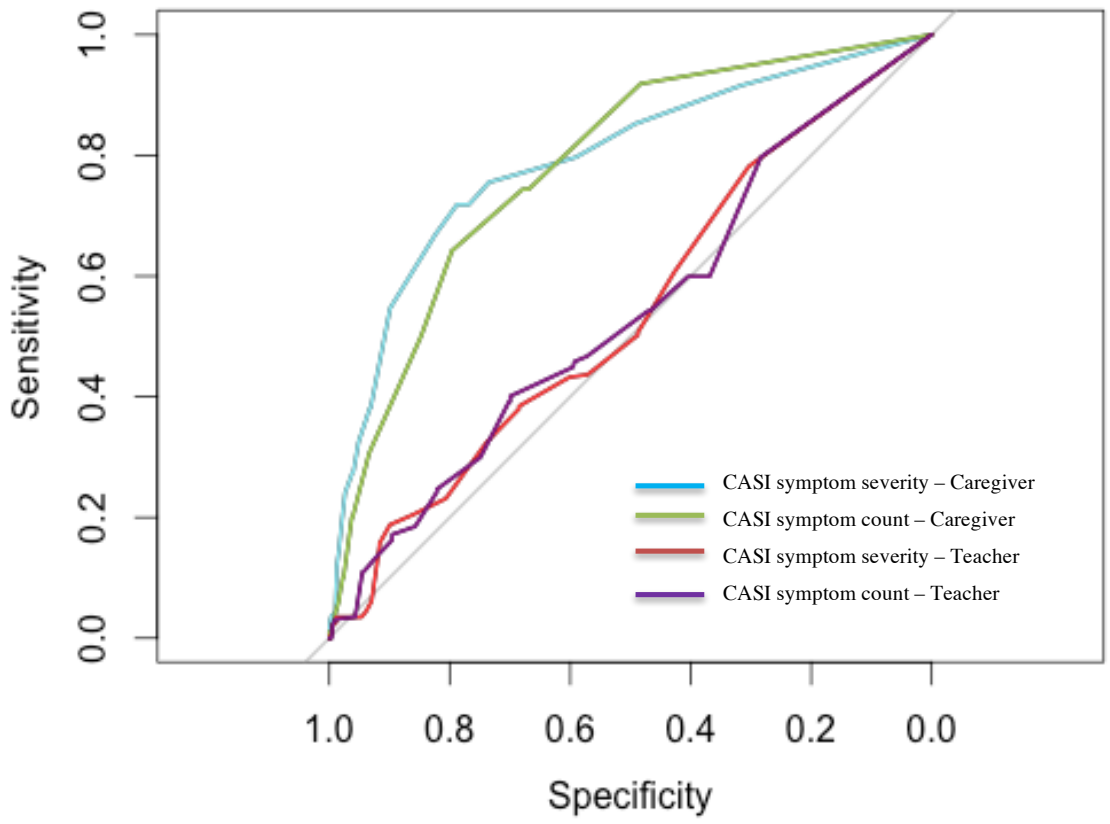

Figure 1. Receiver operating characteristic analyses comparing baseline CASI-4R depression symptom severity and symptom count caregiver reports, and CASI-4R depression symptom severity and symptom count teacher reports.

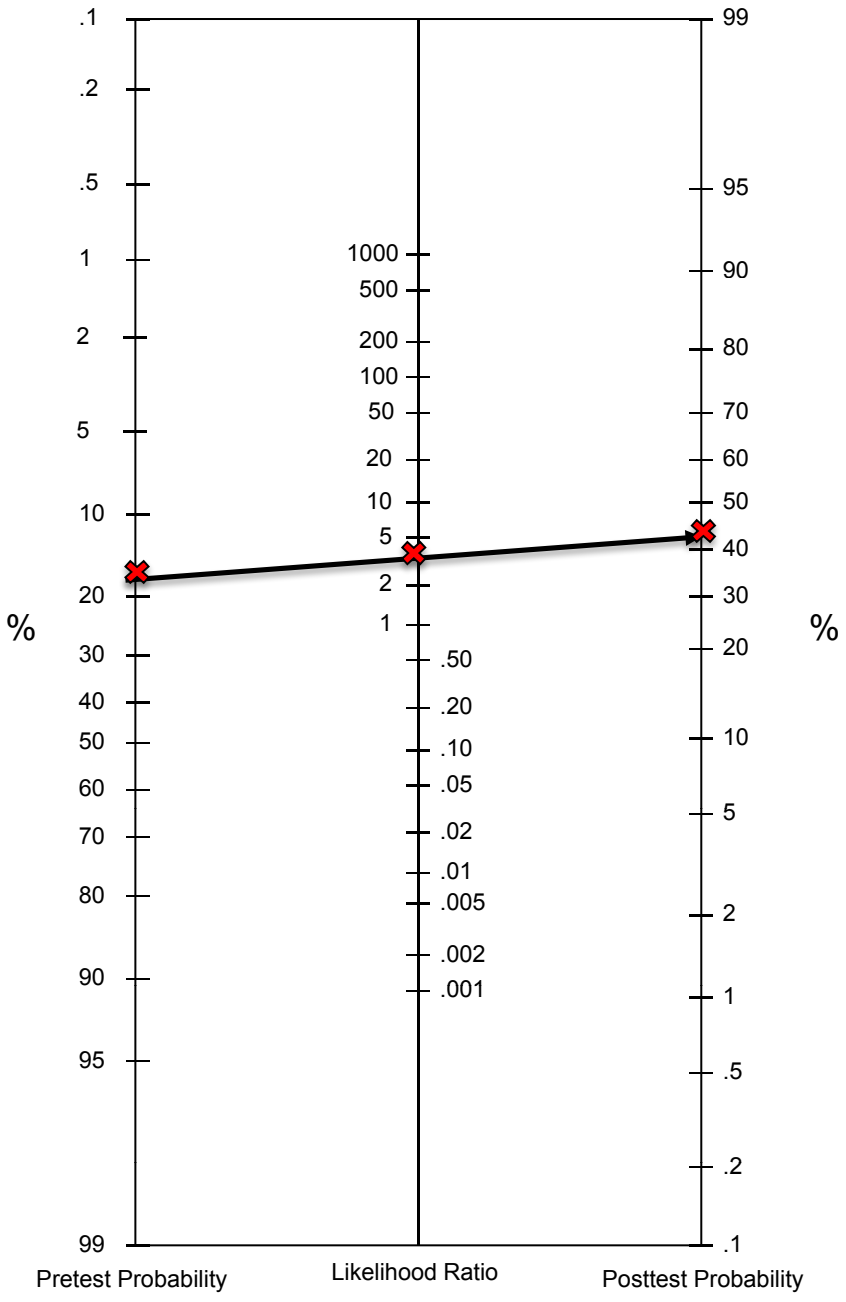

Figure 2. Nomogram example using the CASI-4R depression subscales.

APPENDIX A: DESCRIPTIVE STATISTICS WITH NARROW MOOD DISORDER STATUS

Variable		Any Narrow Mood (*n* = 66)	No Narrow Mood (*n* = 634)	Test Statistic	*p*	Effect Size
Age	*M*	10.08	9.29	$t(86.5)=3.77$	<.001	$d = 0.41$
	SD	1.58	1.92	Levene's $F = 6.55$	.01	
Female		*n* = 24 (36%)	*n* = 202 (32%)	X^2 (1) = 0.55	.46	phi = 0.03
Race (White %)		*n* = 44 (67%)	*n* = 406 (64%)	X^2 (1) = 0.18	.67	phi = 0.02
Axis I Diagnoses at Baseline	*M*	3.33	2.41	$t(698) = 5.74$	<.001	$d = 0.72$
	SD	1.39	1.24	Levene's $F = 0.13$	.72	
CASI-4R Symptom Severity - Caregiver	*M*	8.87	4.57	$t(69.7) = 6.31$	<.001	$d = 0.99$
	SD	5.25	4.05	Levene's $F = 8.78$	.003	
CASI-4R Symptom Severity - Teacher	*M*	5.79	3.62	$t(48.5) = 2.79$	.009	$d = 0.54$
	SD	5.02	3.86	Levene's $F = 5.31$	.02	
CASI-4R Symptom Count - Caregiver	*M*	4.25	2.47	$t(683) = 5.36$	<.001	$d = 0.69$
	SD	2.95	2.48	Levene's $F = 3.06$	.08	
CASI-4R Symptom Count - Teacher	*M*	5.33	3.57	$t(456) = 2.97$	.003	$d = 0.48$
	SD	4.07	3.57	Levene's $F = 1.04$	.31	

Note: Where data points were missing, effect sizes were calculated out of total number of available cases.

APPENDIX B: AUCS FROM ROC ANALYSES FOR NARROW MOOD DISORDER STATUS

		95% Confidence Interval		
Index Test	**Area Under Curve**	**Lower**	**Upper**	**Difference between AUCs**
CASI-4R Symptom Severity - Caregiver	.74***	.68	.80	$p = .08$
CASI-4R Symptom Count - Caregiver	.79***	.75	.83	
CASI-4R Symptom Severity – Teacher	.55	.47	.63	$p < .001$
CASI-4R Symptom Count – Teacher	.56	.50	.63	

Note: *p<.05, **p<.005, ***p<.001, two tailed.

APPENDIX C: DESCRIPTIVE STATISTICS WITH VERY BROAD MOOD DISORDER STATUS

Variable		Any Very Broad Mood (*n* = 269)	No Very Broad Mood (*n* = 431)	Test Statistic	*p*	Effect Size
Age	*M*	9.87	9.05	$t(698)=5.68$	<.001	$d = 0.43$
	SD	1.91	1.84	Levene's $F = 1.49$	.22	
Female		*n* = 103 (38%)	*n* =123 (29%)	X^2 (1) = 7.20	.01	phi = 0.10
Race (White %)		*n* = 181 (67%)	*n* = 269 (62%)	X^2 (1) = 1.71	.19	phi = 0.05
Axis I Diagnoses at Baseline	*M*	3.16	2.07	$t(488.7)=11.52$	<.001	$d = 0.85$
	SD	1.30	1.07	Levene's $F = 14.17$	.001	
CASI-4R Symptom Severity - Caregiver	*M*	7.22	3.59	$t(456.3) = 11.00$	<.001	$d = 0.83$
	SD	4.53	3.60	Levene's $F =14.23$	<.001	
CASI-4R Symptom Severity - Teacher	*M*	4.62	3.31	$t(460) = 3.43$	.001	$d = 0.32$
	SD	4.31	3.77	Levene's $F = 3.24$	.07	
CASI-4R Symptom Count - Caregiver	*M*	3.78	1.93	$t(439.2) = 9.15$	<.001	$d = 0.72$
	SD	2.81	2.13	Levene's $F =33.59$	<.001	
CASI-4R Symptom Count - Teacher	*M*	4.41	3.30	$t(340.2) = 3.11$	.002	$d = 0.30$
	SD	3.90	3.41	Levene's $F =4.51$	.03	

Note: Where data points were missing, effect sizes were calculated out of total number of available cases.

APPENDIX D: AUCS FROM ROC ANALYSES FOR VERY BROAD MOOD DISORDER STATUS

		95% Confidence Interval		
Index Test	**Area Under Curve**	**Lower**	**Upper**	**Difference between AUCs**
CASI-4R Symptom Severity - Caregiver	.79***	.76	.82	$p < .001$
CASI-4R Symptom Count - Caregiver	.78***	.75	.81	
CASI-4R Symptom Severity – Teacher	.55	.51	.60	$p > .05$
CASI-4R Symptom Count – Teacher	.54	.50	.59	

Note: *p<.05, **p<.005, ***p<.001, two tailed.

REFERENCES

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*(2), 213-232.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the Achenbach system of empirically based assessment school-age forms profiles*. Burlington, VT: Aseba.

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Cohen, G. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341-382.

Allgaier, A. K., Fruhe, B., Pietsch, K., Saravo, B., Baethmann, M., & Schulte-Korne, G. (2012). Is the Children's Depression Inventory Short version a valid screening tool in pediatric care? A comparison to its full-length version. *Journal of Psychosomatic Research, 73*(5), 369-374. doi:10.1016/j.jpsychores.2012.08.016

Ambrosini, P. J. (2000). Historical development and present status of the schedule for affective disorders and schizophrenia for school-age children (K-SADS). *Journal of the American Academy of Child and Adolescent Psychiatry, 39*(1), 49-58. doi:10.1097/00004583-200001000-00016

American Psychiatric Association. (1987). Diagnostic and Statistical Manual of Mental Disorders, -revised (DSM-III-R) American Psychiatric Association. *Washington, DC*.

Angold, A. (2002). Diagnostic interviews with parents and children. In M. Rutter & E. Taylor (Eds.), *Child and adolescent psychiatry: Modern approaches* (4th ed., pp. 32-51). Oxford, UK: Blackwell Scientific.

Angold, A., Costello, E. J., & Erkanli, A. (1999). Comorbidity. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 40*(1), 57-87.

Angold, A., Costello, E. J., Messer, S. C., & Pickles, A. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *International Journal of Methods in Psychiatric Research, 5*(4), 237-249.

Bares, C., Andrade, F., Delva, J., Grogan-Kaylor, A., & Kamata, A. (2012). Differential item functioning due to gender between depression and anxiety items among Chilean adolescents. *International Journal of Social Psychiatry, 58*(4), 386-392. doi:10.1177/0020764011400999

Barrera, M., Jr., & Garrison-Jones, C. V. (1988). Properties of the Beck Depression Inventory as a screening instrument for adolescent depression. *Journal of Abnormal Child Psychology, 16*(3), 263-273.

Bayes, T., & Price, R. (1763). 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), 53*, 370-418.

Beck, A. T., & Steer, R. (1993). *Beck Depression Inventory*. San Antonio, TX: Psychological Corporation.

Birmaher, B., Arbelaez, C., & Brent, D. (2002). Course and outcome of child and adolescent major depressive disorder. *Child and Adolescent Psychiatric Clinics of North America, 11*(3), 619-637, x.

Chorpita, B. F., & Daleiden, E. L. (2002). Tripartite dimensions of emotion in a child clinical sample: measurement strategies and implications for clinical utility. *Journal of Consulting and Clinical Psychology, 70*(5), 1150-1160.

Cicchetti, D., & Toth, S. L. (1998). The development of depression in children and adolescents. *American Psychologist, 53*(2), 221-241. doi:10.1037/0003-066X.53.2.221

Clarizio, H. F. (1994). Assessment of depression in children and adolescents by parents, teachers, and peers. In W. M. Reynolds, H. F. Johnston, W. M. Reynolds, & H. F. Johnston (Eds.), *Handbook of depression in children and adolescents*. (pp. 235-248). New York, NY, US: Plenum Press.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100*(3), 316-336.

Cole, D. A., Peeke, L. G., Martin, J. M., Truglio, R., & Seroczynski, A. D. (1998). A longitudinal look at the relation between depression and anxiety in children and adolescents. *Journal of Consulting and Clinical Psychology, 66*(3), 451-460.

Costello, E. J., & Angold, A. (1988). Scales to assess child and adolescent depression: checklists, screens, and nets. *Journal of the American Academy of Child and Adolescent Psychiatry, 27*(6), 726-737. doi:10.1097/00004583-198811000-00011

Crowley, S. L., Worchel, F. F., & Ash, M. J. (1992). Self-report, peer-report, and teacher-report measures of childhood depression: an analysis by item. *Journal of Personality Assessment, 59*(1), 189-203. doi:10.1207/s15327752jpa5901_16

Cyranowski, J. M., Frank, E., Young, E., & Shear, M. K. (2000). Adolescent onset of the gender difference in lifetime rates of major depression: a theoretical model. *Archives of General Psychiatry, 57*(1), 21-27.

Daviss, W. B., Birmaher, B., Melhem, N. A., Axelson, D. A., Michaels, S. M., & Brent, D. A. (2006). Criterion validity of the Mood and Feelings Questionnaire for depressive episodes in clinic and non-clinic subjects. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 47*(9), 927-934. doi:10.1111/j.1469-7610.2006.01646.x

Duax, J. M., Youngstrom, E. A., Calabrese, J. R., & Findling, R. L. (2007). Sex differences in pediatric bipolar disorder. *Journal of Clinical Psychiatry, 68*(10), 1565-1573.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R., & Conover, N. C. (1985). Age differences in the reliability of the psychiatric interview of the child. *Child Development, 56*(1), 265-275.

Eimecke, S. D., Remschmidt, H., & Mattejat, F. (2011). Utility of the Child Behavior Checklist in screening depressive disorders within clinical samples. *Journal of Affective Disorders, 129*(1-3), 191-197. doi:10.1016/j.jad.2010.08.011

Ferdinand, R. F. (2008). Validity of the CBCL/YSR DSM-IV scales Anxiety Problems and Affective Problems. *Journal of Anxiety Disorders, 22*(1), 126-134. doi:10.1016/j.janxdis.2007.01.008

Ferdinand, R. F., Hoogerheide, K. N., van der Ende, J., Visser, J. H., Koot, H. M., Kasius, M. C., & Verhulst, F. C. (2003). The role of the clinician: three-year predictive value of parents', teachers' and clinicians' judgment of childhood psychopathology. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 44*(6), 867-876.

Finch, A. J., Saylor, C. F., Edwards, G. L., & McIntosh, J. A. (1987). Children's Depression Inventory: Reliability over repeated administrations. *Journal of Clinical Child Psychology, 16*(4), 339-341. doi:10.1207/s15374424jccp1604_7

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*(12), 1664-1672. doi:10.4088/JCP.09m05859yel

Ford, T., Goodman, R., & Meltzer, H. (2003). The British Child and Adolescent Mental Health Survey 1999: the prevalence of DSM-IV disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*(10), 1203-1211. doi:10.1097/00004583-200310000-00011

Gadow, K. D., & Sprafkin, J. (2005). Child and adolescent symptom inventory-4R. *Stony Brook, NY: Checkmate Plus*.

Gadow, K. D., & Sprafkin, J. (2015). *The Symptom Inventories: An annotated bibliography*. Stony Brook, NY: Checkmate Plus.

Gadow, K. D., & Sprafkin, J. N. (1998). *Adolescent Symptom Inventory-4*: Checkmate Plus.

Gadow, K. D., & Sprafkin, J. N. (2002). *Child symptom inventory 4: Screening and norms manual*: Checkmate Plus.

Garber, J., & Horowitz, J. L. (2002). Depression in children. In I. H. Gotlib, C. L. Hammen, I. H. Gotlib, & C. L. Hammen (Eds.), *Handbook of depression*. (pp. 510-540). New York, NY, US: Guilford Press.

Garber, J., & Kaminski, K. M. (2000). Laboratory and performance-based measures of depression in children and adolescents. *Journal of Clinical Child Psychology, 29*(4), 509-525. doi:10.1207/S15374424JCCP2904_5

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*(2), 93-100.

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 and Adolescent Psychiatry, 40*(4), 450-455. doi:10.1097/00004583-200104000-00014

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 38*(5), 581-586.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(11), 1337-1345.

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2003). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *International Review of Psychiatry, 15*(1-2), 166-172. doi:10.1080/0954026021000046128

Goodman, R., Meltzer, H., & Bailey, V. (1998). The Strengths and Difficulties Questionnaire: a pilot study on the validity of the self-report version. *European Child and Adolescent Psychiatry, 7*(3), 125-130.

Goodman, R., Renfrew, D., & Mullick, M. (2000). Predicting type of psychiatric disorder from Strengths and Difficulties Questionnaire (SDQ) scores in child mental health clinics in London and Dhaka. *European Child and Adolescent Psychiatry, 9*(2), 129-134.

Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: is small beautiful? *Journal of Abnormal Child Psychology, 27*(1), 17-24.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological Assessment, 12*(1), 19-30.

Hammen, C., & Rudolph, K. D. (2003). Childhood mood disorders. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (pp. 233-278). New York, New York: The Guilford Press.

Hankin, B. L., Abramson, L. Y., Moffitt, T. E., Silva, P. A., McGee, R., & Angell, K. E. (1998). Development of depression from preadolescence to young adulthood: emerging gender

differences in a 10-year longitudinal study. *Journal of Abnormal Psychology, 107*(1), 128-140.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*(3), 839-843. doi:10.1148/radiology.148.3.6878708

Harrington, R., Rutter, M., & Fombonne, E. (1996). Developmental pathways in depression: Multiple meanings, antecedents, and endpoints. *Development and Psychopathology, 8*(4), 601-616. doi:10.1017/S095457940000732X

Hart, E. L., & Lahey, B. B. (1999). General child behavior rating scales. In D. Shaffer, C. P. Lucas, J. E. Richters, D. Shaffer, C. P. Lucas, & J. E. Richters (Eds.), *Diagnostic assessment in child and adolescent psychopathology*. (pp. 65-87). New York, NY, US: Guilford Press.

He, J. P., Burstein, M., Schmitz, A., & Merikangas, K. R. (2013). The Strengths and Difficulties Questionnaire (SDQ): the factor structure and scale validation in U.S. adolescents. *Journal of Abnormal Child Psychology, 41*(4), 583-595. doi:10.1007/s10802-012-9696-6

Hilt, L., & Nolen-Hoeksema, S. (2009). The emergence of gender differences in depression in adolescence. In S. Nolen-Hoeksema & L. M. Hilt (Eds.), *Handbook of depression in adolescents* (pp. 111-135). New York: Routledge.

Horwitz, S. M., Demeter, C. A., Pagano, M. E., Youngstrom, E. A., Fristad, M. A., Arnold, L. E., . . . Findling, R. L. (2010). Longitudinal Assessment of Manic Symptoms (LAMS) study: background, design, and initial screening results. *Journal of Clinical Psychiatry, 71*(11), 1511-1517. doi:10.4088/JCP.09m05835yel

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *American Journal of Psychiatry, 167*(7), 748-751. doi:10.1176/appi.ajp.2010.09091379

Jacques, H. A., & Mash, E. J. (2004). A test of the tripartite model of anxiety and depression in elementary and high school boys and girls. *Journal of Abnormal Child Psychology, 32*(1), 13-25.

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*(2), 121-129. doi:10.1037/a0022506

Jensen, B. J., & Haynes, S. N. (1986). Self-Report Questionnaires and Inventories. In A. S. Ciminero, C. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 150-175). New York: Wiley.

Jensen, P. S., Rubio-Stipec, M., Canino, G., Bird, H. R., Dulcan, M. K., Schwab-Stone, M. E., & Lahey, B. B. (1999). Parent and child contributions to diagnosis of mental disorder: are both informants always necessary? *Journal of the American Academy of Child and Adolescent Psychiatry, 38*(12), 1569-1579. doi:10.1097/00004583-199912000-00019

Johnson, S., Hollis, C., Marlow, N., Simms, V., & Wolke, D. (2014). Screening for childhood mental health disorders using the Strengths and Difficulties Questionnaire: the validity of multi-informant reports. *Developmental Medicine and Child Neurology, 56*(5), 453-459. doi:10.1111/dmcn.12360

Kaufman, J., Birmaher, B., Brent, D., & Rao, U. (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*(7), 980-988.

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 and Adolescent Psychiatry, 36*(7), 980-988. doi:10.1097/00004583-199707000-00021

Kessler, R. C., Avenevoli, S., Costello, E. J., Georgiades, K., Green, J. G., Gruber, M. J., . . . Merikangas, K. R. (2012). Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. *Archives of General Psychiatry, 69*(4), 372-380. doi:10.1001/archgenpsychiatry.2011.160

Kessler, R. C., Avenevoli, S., & Ries Merikangas, K. (2001). Mood disorders in children and adolescents: an epidemiologic perspective. *Biological Psychiatry, 49*(12), 1002-1014.

Klein, D. N., Dougherty, L. R., & Olino, T. M. (2005). Toward guidelines for evidence-based assessment of depression in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 412-432. doi:10.1207/s15374424jccp3403_3

Klein, D. N., Riso, L. P., & Anderson, R. L. (1993). DSM-III-R dysthymia: antecedents and underlying assumptions. *Progress in Experimental Personality and Psychopathology Research, 16*, 222-253.

Kovacs, M. (1986). A developmental perspective on methods and measures in the assessment of depressive disorders: The clinical interview. In M. Rutter, C. E. Izard, & P. B. Read (Eds.), *Depression in young people: Developmental and clinical perspectives* (pp. 435-465). New York: Guilford.

Kovacs, M. (1992). *Children's Depression Inventory: Manual*. North Tonawanda, NY: Multi-Health Systems.

Kovacs, M. (1996). 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, 35*(6), 705-715. doi:10.1097/00004583-199606000-00010

Kovacs, M. (2003). *Children's Depression Inventory: Technical Manual Update*. North Tonawanda, NY: Multi-Health Systems.

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. New York: Academic Press.

Lengua, L. J., Sadowski, C. A., Friedrich, W. N., & Fisher, J. (2001). Rationally and empirically derived dimensions of children's symptomatology: expert ratings and confirmatory factor analyses of the CBCL. *Journal of Consulting and Clinical Psychology, 69*(4), 683-698.

Lewinsohn, P. M., & Essau, C. A. (2002). Depression in adolescents. In I. H. Gotlib, C. L. Hammen, I. H. Gotlib, & C. L. Hammen (Eds.), *Handbook of depression*. (pp. 541-559). New York, NY, US: Guilford Press.

Lewinsohn, P. M., Hops, H., Roberts, R. E., Seeley, J. R., & Andrews, J. A. (1993). Adolescent psychopathology: I. Prevalence and incidence of depression and other DSM-III-R disorders in high school students. *Journal of Abnormal Psychology, 102*(1), 133-144.

Lewinsohn, P. M., Rohde, P., Seeley, J. R., Klein, D. N., & Gotlib, I. H. (2003). Psychosocial functioning of young adults who have experienced and recovered from major depressive disorder during adolescence. *Journal of Abnormal Psychology, 112*(3), 353-363.

Little, R. J., & Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine, 22*(9), 1589-1599.

Lonigan, C. J., Phillips, B. M., & Hooe, E. S. (2003). Relations of positive and negative affectivity to anxiety and depression in children: evidence from a latent variable longitudinal study. *Journal of Consulting and Clinical Psychology, 71*(3), 465-481.

Masip, A. F., Amador-Campos, J. A., Gómez-Benito, J., & Gándara, V. d. B. (2010). Psychometric properties of the Children's Depression Inventory in community and clinical sample. *The Spanish Journal of Psychology, 13*(02), 990-999.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*, 215-241. doi:10.1146/annurev.psych.50.1.215

Myers, K., & Winters, N. C. (2002). Ten-year review of rating scales. II: Scales for internalizing disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 41*(6), 634-659. doi:10.1097/00004583-200206000-00004

Najman, J. M., Williams, G. M., Nikles, J., Spence, S., Bor, W., O'Callaghan, M., . . . Andersen, M. J. (2000). Mothers' mental illness and child behavior problems: cause-effect

association or observation bias? *Journal of the American Academy of Child and Adolescent Psychiatry, 39*(5), 592-602. doi:10.1097/00004583-200005000-00013

Nezu, A. M., Ronan, G. F., Meadows, E. A., & McClure, K. S. (2000). *Practitioner's guide to empirically based measures of depression*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Nolan, T. M., Bond, L., Adler, R., Littlefield, L., Birleson, P., Marriage, K., . . . Tonge, B. J. (1996). Child Behaviour Checklist classification of behaviour disorder. *Journal of Paediatrics and Child Health, 32*(5), 405-411.

Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology, 109*(3), 504-511.

Rao, U., Ryan, N. D., Birmaher, B., Dahl, R. E., Williamson, D. E., Kaufman, J., . . . Nelson, B. (1995). Unipolar depression in adolescents: clinical outcome in adulthood. *Journal of the American Academy of Child and Adolescent Psychiatry, 34*(5), 566-578. doi:10.1097/00004583-199505000-00009

Reich, W., Shayka, T., & Taibelson, C. (1991). *The Diagnostic Interview for Children and Adolescents - Revised* St. Louis, MO: Washington University.

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*(3), 169-184. doi:10.1002/mpr.289

Rey, J. M., Schrader, E., & Morris-Yates, A. (1992). Parent-child agreement on children's behaviours reported by the Child Behaviour Checklist (CBCL). *Journal of Adolescence, 15*(3), 219-230.

Richters, J., & Pellegrini, D. S. (1989). Depressed mothers' judgments about their children: An examination of the depression-distortion hypothesis. *Child Development, 60*(5), 1068-1075. doi:10.2307/1130780

Richters, J. E. (1992). Depressed mothers as informants about their children: a critical review of the evidence for distortion. *Psychological Bulletin, 112*(3), 485-499.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Muller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 77.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 1.

Rohde, P. (2009). Comorbidities with adolescent depression. In S. Nolen-Hoeksema, L. M. Hilt, S. Nolen-Hoeksema, & L. M. Hilt (Eds.), *Handbook of depression in adolescents*. (pp. 139-177). New York, NY, US: Routledge/Taylor & Francis Group.

Sanislow, C. A., Pine, D. S., Quinn, K. J., Kozak, M. J., Garvey, M. A., Heinssen, R. K., . . . Cuthbert, B. N. (2010). Developing constructs for psychopathology research: research domain criteria. *Journal of Abnormal Psychology, 119*(4), 631-639. doi:10.1037/a0020909

Saylor, C. F., Finch, A. J., Spirito, A., & Bennett, B. (1984). The Children's Depression Inventory: A systematic evaluation of psychometric properties. *Journal of Consulting and Clinical Psychology, 52*(6), 955-967. doi:10.1037/0022-006X.52.6.955

Shih, J. H., Eberhart, N. K., Hammen, C. L., & Brennan, P. A. (2006). Differential exposure and reactivity to interpersonal stress predict sex differences in adolescent depression. *Journal of Clinical Child and Adolescent Psychology, 35*(1), 103-115. doi:10.1207/s15374424jccp3501_9

Silverman, W. K., Kurtines, W. M., Ginsburg, G. S., Weems, C. F., Rabian, B., & Serafini, L. T. (1999). Contingency management, self-control, and education support in the treatment of childhood phobic disorders: a randomized clinical trial. *Journal of Consulting and Clinical Psychology, 67*(5), 675-687.

Smucker, M. R., Craighead, W. E., Craighead, L. W., & Green, B. J. (1986). Normative and reliability data for the Children's Depression Inventory. *Journal of Abnormal Child Psychology, 14*(1), 25-39.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24*(5), 399-411. doi:10.1016/0010-440X(83)90032-9

Sprafkin, J., Gadow, K. D., Salisbury, H., Schneider, J., & Loney, J. (2002). 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, 31*(4), 513-524. doi:10.1207/S15374424JCCP3104_10

Stockings, E., Degenhardt, L., Lee, Y. Y., Mihalopoulos, C., Liu, A., Hobbs, M., & Patton, G. (2015). 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, 174*, 447-463. doi:10.1016/j.jad.2014.11.061

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

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293. doi:10.1126/science.3287615

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest, 1*(1), 1-26. doi:10.1111/1529-1006.001

Thapar, A., & McGuffin, P. (1998). Validity of the shortened Mood and Feelings Questionnaire in a community sample of children and adolescents: a preliminary research note. *Psychiatry Research, 81*(2), 259-268.

Timbremont, B., Braet, C., & Dreessen, L. (2004). Assessing depression in youth: relation between the Children's Depression Inventory and a structured interview. *Journal of Clinical Child and Adolescent Psychology, 33*(1), 149-157. doi:10.1207/S15374424JCCP3301_14

van Beek, Y., Hessen, D. J., Hutteman, R., Verhulp, E. E., & van Leuven, M. (2012). Age and gender differences in depression across adolescence: Real or 'bias'? *Journal of Child Psychology and Psychiatry, 53*(9), 973-985. doi:10.1111/j.1469-7610.2012.02553.x

Van Meter, A., You, D. S., Halverson, T., Youngstrom, E. A., Birmaher, B., Fristad, M., . . . LAMS Group. (under revision). Diagnostic Efficiency of Caregiver Report on the SCARED for Identifying Youth Anxiety Disorders in Outpatient Settings

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics, 56*(4), 1134-1138.

Venkatraman, E. S., & Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika, 83*(4), 835-848. doi:10.1093/biomet/83.4.835

Verhulst, F. C., Dekker, M. C., & van der Ende, J. (1997). Parent, teacher and self-reports as predictors of signs of disturbance in adolescents: whose information carries the most weight? *Acta Psychiatrica Scandinavica, 96*(1), 75-81.

Waslick, B. D., Kandel, R., & Kakouros, A. (2002). Depression in children and adolescents: An overview. In D. Shaffer, B. D. Waslick, D. Shaffer, & B. D. Waslick (Eds.), *The many faces of depression in children and adolescents*. (pp. 1-36). Arlington, VA, US: American Psychiatric Publishing, Inc.

Watson, D., Clark, L. A., Weber, K., Assenheimer, J. S., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *Journal of Abnormal Psychology, 104*(1), 15-25.

Weiss, B., & Garber, J. (2003). Developmental differences in the phenomenology of depression. *Development and Psychopathology, 15*(2), 403-430. doi:10.1017/S0954579403000221

Wood, A., Kroll, L., Moore, A., & Harrington, R. (1995). Properties of the mood and feelings questionnaire in adolescent psychiatric outpatients: a research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 36*(2), 327-334.

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. doi:10.1093/jpepsy/jst062

Youngstrom, E. A., Choukas-Bradley, S., Calhoun, C. D., & Jensen-Doss, A. (2014). Clinical guide to the Evidence-Based Assessment approach to diagnosis and treatment. *Cognitive and Behavioral Practice*. doi:10.1016/j.cbpra.2013.12.005

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*(1), 20-35. doi:10.1016/j.cbpra.2013.12.005

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*(3), 231-245.

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 and Adolescent Psychiatry, 43*(7), 847-858. doi:10.1097/01.chi.0000125091.35109.1e

Youngstrom, E. A., Frazier, T. W., Demeter, C., Calabrese, J. R., & Findling, R. L. (2008). Developing a 10-item mania scale from the Parent General Behavior Inventory for children and adolescents. *Journal of Clinical Psychiatry, 69*(5), 831-839.

Youngstrom, E. A., Izard, C., & Ackerman, B. (1999). Dysphoria-related bias in maternal ratings of children. *Journal of Consulting and Clinical Psychology, 67*(6), 905-916.

Youngstrom, E. A., 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 Disord, 7*(6), 507-517. doi:10.1111/j.1399-5618.2005.00269.x

Youngstrom, E. A., Meyers, O., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006). Comparing the effects of sampling designs on the diagnostic accuracy of eight promising screening algorithms for pediatric bipolar disorder. *Biological Psychiatry, 60*(9), 1013-1019. doi:10.1016/j.biopsych.2006.06.023