UNIVERSAL EMOTIONAL AND BEHAVIORAL SCREENING FOR CHILDREN AND ADOLESCENTS: ARE MULTIPLE GATES AND INFORMANTS WORTH IT?

by

MEGHAN C. VANDEVENTER

(Under the Direction of Randy W. Kamphaus)

ABSTRACT

Emotional and behavioral adjustment is undeniably intertwined with children's general physical health, academic achievement, and adaptation throughout their lives. Furthermore, child emotional and behavioral disorders tend to persist into adulthood. Early identification and intervention of youth with emotional and behavioral problems can help to minimize the long-term detrimental effects of mental disorders. Unfortunately, our current mental health care system is not succeeding in identifying those children in need of services. Universal emotional and behavioral screening is an efficient way to assess all children and identify those children at-risk for specific emotional and behavioral problems, allowing clinicians to act early so as to reduce risk, prevent the onset, or minimize the effects of a disorder.

Researchers have called for a multiple gate screening procedure which begins by screening an entire population for emotional and behavioral difficulties. Those children identified by the screening instrument as at-risk for emotional and behavioral problems are then assessed again using a different, often more thorough, assessment tool such as a full behavior rating scale in order to increase accuracy of identification. When

developing such a procedure, one must evaluate the utility of multiple gates as well as which informants should constitute each gate. The current study assessed the known-groups validity of newly developed BASC-2 screeners as well as examined the issues of gates and informants when implementing a universal screening procedure.

The BASC-2 screeners appear to be promising as first gate screening measures. Additionally, parents were found to do a better job than teachers as first gate screener informants. Adding a comprehensive behavior rating scale as a second gate significantly improved classification accuracy and utilizing a different informant at the second gate appeared to improve classification accuracy further. Lastly, when implementing a two informant, two gate screening procedure, a screener is a valid option as the second gate in place of a longer behavior rating scale such as the full BASC-2. More research must still be done in order to ensure that sound science guides the increasingly popular practice of screening children for behavioral and emotional problems.

INDEX WORDS: Screening, Children, Adolescents, Emotional and Behavioral Adjustment, Multiple gates, Multiple Informants

UNIVERSAL EMOTIONAL AND BEHAVIORAL SCREENING FOR CHILDREN AND ADOLESCENTS: ARE MULTIPLE GATES AND INFORMANTS WORTH IT?

by

MEGHAN C. VANDEVENTER

B.A., University of Notre Dame, 2003

M.Ed., University of Georgia, 2005

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

UNIVERSAL EMOTIONAL AND BEHAVIORAL SCREENING FOR CHILDREN AND ADOLESCENTS: ARE MULTIPLE GATES AND INFORMANTS WORTH IT?

by

MEGHAN C. VANDEVENTER

Major Professor:	Randy W. Kamphaus, Ph.D.
Committee:	Deborah Bandalos, Ph.D. Jonathan M. Campbell, Ph.D. Roy P. Martin, Ph.D.

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2008

DEDICATION

This work is dedicated to my parents, without whom I would be lost in so many ways. Thank you for your constant love and support, for your unfailing confidence in me and my abilities, and for always putting things into perspective. Thank you for giving me wings and for being the wind beneath them.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor and mentor, Randy Kamphaus. The completion of this work would not have been possible without his words of wisdom, constant support, and patient guidance. I will be forever grateful for his humor, complete dedication to his students, and ability to continually challenge me to go one step beyond what I ever thought possible. I would also like to thank my committee members, Deborah Bandalos, Jonathan Campbell, and Roy Martin for their unlimited help and support. Lastly, thank you to my family and friends for keeping me sane and making me laugh throughout this journey.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND REVIEW OF THE LITERATURE

The present state of child and adolescent mental health in the United States has become an area of major concern across the highest levels of government, including the President of the United States and members of both the House of Representatives and the Senate. On September 18 and 19, 2000, the *Surgeon General's Conference on Children's Mental Health: Developing a National Action Agenda* was held in Washington D.C. to address these concerns and develop specific recommendations for a National Action Agenda on Children's Mental Health. In 2001, the American Psychological Association created a *Working Group on Children's Mental Health* (Tolan & Dodge, 2005). In 2003, President George Bush established the *New Freedom Commission*, in response to the 1999 White House Conference on Mental Health, to identify policies that could be implemented by federal, state and local governments to address, among other things, the shortcomings found in the current identification and treatment practices for children with emotional disturbances.

In May 2005, the Campaign for Mental Health Reform addressed the United States Senate and House of Representatives, identifying the current state of child and adolescent mental health services as a "public health crisis." In order to be convinced of the idea that we are in the midst of a "public health crisis," one must understand the importance of mental health to our overall well-being as well as the inadequacy of our current mental health system of care and prevention.

Importance of Child Mental Health

Evidence abounds that children's emotional and behavioral adjustment is intertwined with their general physical health and academic achievement, and also has been linked to successful adaptation throughout their lives (Masten & Coatsworth, 1998). Mental disorder ranks first among illnesses that cause disability in the United States, Canada, and Western Europe (New Freedom Commission on Mental Health, 2003). In the United States, mental illness accounts for more than 15% of the overall burden of disease (United States Department of Health and Human Services, 1999). Suicide has also been found to be a major problem worldwide, emerging as the third leading cause of death in youth ages 15 to 24. Over 90% of children and adolescents who commit suicide have at least one mental disorder (Campaign for Mental Health Reform, 2005).

Research has also demonstrated that the effects of child emotional and behavioral disorders, as well as the disorders themselves, tend to persist into adulthood with 74% of 21 year olds with mental disorders having had prior mental health problems (United States Public Health Service, 2000; Aronen, Teerikangas, & Kurkela, 1999). Children with emotional and behavioral problems are more likely to drop out of school, abuse substances, be involved in the juvenile justice system, and commit suicide. Strikingly, approximately 50% of students ages 14 and older with a mental disorder will drop out of school; only 42% of those who remain will graduate with a diploma (United States Public Health Service, 2000). Additionally, 65% of boys and 75% of girls in juvenile detention centers have a least one mental disorder (Campaign for Mental Health Reform, 2005).

The cost to society is high not only in human, but in financial terms. Between 5% and 10% of the total cost and morbidity burden due to disease is accounted for by mental

disorders (Jenkins, 1998). In the United States, the indirect cost of mental illness is about $79 billion annually. This figure includes costs due to loss of productivity, incarceration, as well as treatment. When children with untreated emotional and behavioral disorders become adults, they continue to utilize more health care services and incur much higher health care costs than other adults. States spend nearly $1 billion per year on medical costs associated with completed suicides and suicide attempts by youth (Campaign for Mental Health Reform, 2005). Cohen (1998) found that diverting one high risk child from developing serious conduct problems may result in a savings of nearly $2 million to society.

Research has demonstrated that early identification and intervention for youth with emotional and behavioral problems can help to minimize the long-term detriment of mental disorders as well as reduce the overall healthcare burden and costs (Aos, Lieb, Mayfield, Miller, & Pennucci, 2004; Campaign for Mental Health Reform, 2005). Early emotional and behavioral difficulties, including subsyndromal symptomatology, can lead to a pattern of adjustment problems that may be transient or long standing, depending on the services provided and the timing of these services. The longer a child's emotional and behavioral problems go unidentified, the more stable his or her maladaptive trajectory is likely to be (Gottlieb, 1991).

Younger children exhibit more plasticity and malleability both behaviorally and neurodevelopmentally thus making their maladaptive behaviors easier to modify (Hirshfield-Becker & Biederman, 2002). Early identification and intervention also catches developing problems before they become more severe or expand into numerous co-occurring disorders. Untreated emotional and behavioral problems during this crucial

time tend to persist into later childhood and adulthood, interfering with the development of critical emotional and cognitive skills, escalating in severity, and leading to a downward spiral of school failure, unemployment, substance abuse, and poverty (United States Public Health Service, 2000; McGoey, Eckert, & Dupaul, 2002; Hirshfield-Becker & Biederman, 2002). Through early identification and treatment, we may prevent negative lifelong outcomes. Thus childhood is an essential time to identify and prevent mental disorders as well as promote healthy development.

Current Mental Health Care System

The current mental health care system has little chance of succeeding as it fails at the outset by not identifying children in need of services. Recent research indicates that approximately 1 in 5 children have a diagnosable mental disorder; furthermore, 10-13% of preschoolers, ages 1-6, have emotional/behavioral disorders (Campaign for Mental Health Reform, 2005; Friedman, Katz-Leavy, Manderscheild, & Sondheimer, 1996). Thus, many of these problems begin early and many are ignored beginning in preschool. For example, Fantuzzo, Bulotsky, McDermott, Mosca, and Lutz (2003) found that Head Start staff under-identified children with behavioral or emotional problems as a group and, those children with the highest risk for poor academic readiness were most likely to be unidentified and untreated.

In general, only 15 to 20% of children with emotional and behavioral problems receive any type of mental health services in a given year (Ringel & Sturm, 2001; United States Public Health Service, 2000). Jenkins (1998) estimated that mental health specialists are able to meet the need of only 10% of all children with emotional and behavioral problems. Generally, the children who do receive services are those with the

most fully developed and severe mental disorders; however, mental health needs extend beyond diagnosable disorders. Children who exhibit signs of risk or subsyndromal symptomatology need access to interventions as well, in order to prevent the development of more serious disorders. Thus, although the most effective way to maximize the likelihood of positive treatment outcomes is to identify and treat children early, they must often wait until their problems are "serious enough" before they can receive services, often referred to as the "Wait-to-Fail" approach. Alarmingly, the "National Comorbidity Survey Replication" study found that the median lag between the onset of a mental disorder and the start of treatment is about 10 years. Disorders emerging in childhood have the longest delays in treatment perhaps due to reliance on parents or other adults as informants (National Mental Health Association, 2005). A critical gap exists between those who need mental health services and those who actually receive them. This unmet need for services remains as high as it was twenty years ago (United States Public Health Service, 2000).

Settings for Identification

Primary care and school settings appear to be two of the most important systems for the potential early detection of emotional and behavioral problems in children and adolescents. The majority of children are seen in these settings thus providing the opportunity to reach large numbers of youth. A number of challenges exist, however, that contribute to the failure of these settings to identify and treat children with emotional and behavioral problems.

In primary care settings, we first must address the problem of access to medical care. Approximately 16% of the United States adult population does not have health

insurance (United States Department of Health and Human Services, 1999). Additionally, a lack of parity exists in insurance coverage of general health versus mental health services. In fact, studies show that the gap in insurance coverage between mental health and other health services has been widening (United States Department of Health and Human Services, 1999). Therefore, even those who have private health insurance may find it difficult to finance mental health services. Secondly, the stigma of mental disorders still exists in our society and deters Americans from seeking care. As President Bush stated, "…Americans must understand and send this message: mental disability is not a scandal – it is an illness. And like physical illness, it is treatable, especially when the treatment comes early" (New Freedom Commission on Mental Health, 2003).

Furthermore, parents are often unaware that they can discuss mental health issues with primary care physicians, assuming that primary care physicians are only concerned with physical health. Arcia and Fernandez (2003) found that Latino mothers were also more likely to seek services for their children when they exhibited behaviors such as hyperactivity and aggression rather than internalizing symptomatology such as anxiety. Lastly, there tends to be poor recognition of mental illness by physicians due to time limitations, limited training in mental health issues, and a focus on the central task of assessing physical health. The average visit with a primary care physician is only between 11 and 15 minutes (United States Department of Health and Human Services, 1999). Although an increasing number of emotional and behavioral problems (15-30%) are being identified by primary care providers, rates of recognition (48-57%) are still low and connections to mental health specialists are unlikely (United States Department of Health and Human Services, 1999).

In schools, children with emotional and behavioral needs are usually identified only after their problems cannot be managed by their regular classroom teacher. Most identification is done through teacher-initiated referral for evaluation – an idiosyncratic, externalizing behavior problem- focused method that allows many children with emotional and behavioral, especially internalizing, problems to fall through the cracks. Lloyd, Kauffman, Landrum, and Roe (1991) found that general education teachers are involved in 79% of all school referrals. Several studies have indicated that those students who are referred usually exhibit externalizing behavior problems that are highly disruptive and aversive to both teachers and peers (Grosenick, 1981; Noel, 1982) rather than internalizing problems such as depression, shyness, phobias, or social avoidance (Walker, Severson, Stiller, Williams, Haring, Shinn, & Todis, 1988). Additionally, Lloyd and colleagues (1991) found that 69% of teacher referrals were for males. Even if a child is referred, a series of parent conferences, discipline referrals, and trial interventions in the regular classroom often precede an actual referral for mental health services. In California, Forness (United States Public Health Service, 2000) found that schools are not adequately identifying children in need of mental health services nor identifying them soon enough (United States Department of Health and Human Services, 1999).

A recent review of the literature by Jamieson and Romer (2005) called for a national effort to reorganize current mental health identification processes through the following observation:

> Because it is clear that early detection and referral for treatment should be a high national priority, it is disappointing to learn from research conducted as part of the commissions (referring to expert panels created as part of the Adolescent Mental

> Health Initiative of the Annenberg Foundation Trust)….that the primary care system and schools are inadequately prepared to meet this challenge…..As a result, schools do not intervene until illnesses progress and come to the attention of staff. (p. 619).

Improving Service Delivery

The current mental health service delivery system could be improved by adopting a more unified system of health care where children are served in one central location rather than the fragmented system that exists today in which children are constantly under-identified for services. Many researchers and clinicians (Huang, Stroul, Friedman, Mrazek, Friesen, Pires, & Mayberg, 2005; Tolan & Dodge, 2005) have begun to advocate for a "systems of care" approach to child mental health service delivery that embraces principles such as "wrapping services around the child rather than requiring the child to conform to the provider's culture and construal of care" and "including all service providers in a unified plan" (Tolan & Dodge, 2005, p. 608). This approach basically states that mental health services should be provided in settings where children are already seen, such as primary care and educational settings, thereby increasing the accessibility of these services for parents and children.

More schools are beginning to provide school-based health centers which offer a one-stop source for medical, psychological, and preventative care. Schools are "where the children are" and therefore provide an optimal setting for service delivery. According to the National Assembly on School-Based Health Care (NASBHC), the number of such centers has increased from 120 sites in 1988 to more than 1,500 in 2005 (Martin, 2005). This involvement of the education system in child mental health is consistent with current

legislative demands such as No Child Left Behind (2001) in that children's emotional and behavioral problems have been found to have a significant adverse effect on academic achievement (Gutman, Sameroff, & Cole, 2003; Huang et al., 2005; Jimerson, Egeland, & Teo, 1999; McEvoy & Welker, 2000; Rapport, Denney, Chung, & Hustace, 2001).

Universal Screening of Emotional and Behavioral Adjustment

A common thread found in all governmental and non-governmental action plans related to children's mental health is the need for universal screening and early identification of children and adolescents for emotional and behavioral problems. For example, the *Report of the Surgeon General's Conference on Children's Mental Health* calls for screening and early identification of children within key service systems as well as the development of "a universal measurement system across all major service sectors that is age-appropriate, culturally-competent, and gender sensitive to (i) identify children, including those with special healthcare needs, who may need mental health services; (ii) track child progress during treatment; and (iii) measure treatment outcomes for individual patients" (United States Public Health Service, 2000). Through universal screening of emotional and behavioral adjustment we can work to reduce risk, prevent onset, and intervene early so as to improve outcomes significantly.

The principles of prevention, early detection, and universal care were first applied to the field of infectious disease through the use of vaccinations, water safety, and other forms of public hygiene practices. At birth, infants are screened for a number of genetic diseases and children are routinely screened for hearing, vision, and scoliosis at school and during pediatrician visits. These universal screening and care practices are now standard practice for physical health concerns and have been largely successful in

minimizing the negative effects and, in come cases, eliminating certain diseases from the population.

In 1968, the World Health Organization (WHO; Wilson & Jungner, 1968) provided guidelines for effective health screening that provide a framework for thinking through the process of mental health screening:

1. The condition should represent an important health problem that carries with it notable morbidity and mortality.
2. Screening programs must be cost-effective, that is, the incidence/significance of the disorder must be sufficient to justify the costs of screening
3. Effective methods of treatment must be available for the disorder.
4. The test(s) for the disorder should be reliable and valid so that detection errors (i.e., false positives or false negatives) are minimized.
5. The test(s) should have high cost-benefit, that is the time, effort, and personal inconvenience to the patient associated with taking the test should be substantially outweighed by its potential benefits.
6. The condition should be characterized by an asymptomatic or benign period, during which detection will significantly reduce morbidity and/or mortality.
7. Treatment administered during the asymptomatic phase should demonstrate significantly greater efficacy than that dispensed during the symptomatic phase.

Generally, it appears that mental health disorders meet these criteria. As presented previously, mental health disorders are important health problems with known morbidity, mortality, and costs to society (Campaign for Mental Health Reform, 2005; United States Public Health Service, 2000). Evidence-based medications as well as

psychosocial interventions exist that have been found to effectively treat most mental health disorders (New Freedom Commission on Mental Health, 2003). A number of effective or promising treatments exist for many mental disorders in children including cognitive-behavioral therapy and selective serotonin reuptake inhibitors for depression (Kaslow & Thompson, 1998), parent training and multisystemic therapy for conduct disorder (Brestan & Eyberg, 1998), and psychostimulants and behavioral training of teachers for attention-deficit/hyperactivity disorder (Pelham, Wheeler, & Chronis, 1998). Research has also shown that identifying and treating children early, before their emotional and behavior problems are diagnosable, can minimize the long-term detriment of mental disorders as well as reduce the overall healthcare burden and costs (Aos, Lieb, Mayfield, Miller, & Pennucci, 2004; Campaign for Mental Health Reform, 2005).

Jones, Dodge, Foster, Nix, and the Conduct Problems Prevention Research Group (2002) found that an inexpensive screening in kindergarten predicted children involved in mental health, special education, or juvenile justice services six years later; however, emotional and behavioral screening is currently minimal to nonexistent throughout the U.S. health care system. Only 2% of schools screen for emotional and behavioral problems (Romer & McIntosh, 2005) and routine developmental and psychosocial assessments of young children using standardized instruments in pediatric settings are just as rare (United States Public Health Service, 2000). As Huang stated, “By avoiding this, we are perpetuating stigma and failing to normalize mental health and recognize it as a critical part of overall health and well-being” (Campaign for Mental Health Reform, 2005).

In addition, universal emotional and behavioral screening at school is an efficient, quick way to assess all children, identify those children at-risk for specific illnesses and disorders, and act early so as to reduce risk, prevent the onset, or minimize the effect of the disorder. Through universal screening, we have the potential to not only identify a greater proportion of individuals with emotional and behavioral problems, but to do so at an earlier stage thereby reducing the severity and long - term impact of the disorder. Moreover, universal emotional and behavioral screening can save monetary and time resources by minimizing the number of unnecessary diagnostic tests as well as reducing length and need for treatment and hospitalizations. However, the success of early intervention depends on the accuracy and utility of the method used to identify high risk children. More research must be done in order to develop screening instruments and programs and determine whether these programs are valid, cost-effective, and adequately beneficial.

Evaluating Screening Instruments

When evaluating a screening instrument, researchers first must evaluate the psychometric properties of the measure including norm adequacy, reliability, and validity. Validity, as defined by Messick (1995), is "an integrated judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment." Messick (1995) has concluded that construct validity is the most important type of validity and, in actuality, subsumes all other types of validity including content and predictive validity. In assessing the validity of a test, the goal is not to conclude whether the test is valid or not but rather to state as definitively as possible the degree of

validation. Thus construct validity can be viewed as an accumulation of evidence over time and is not unlike the general scientific procedures for developing and confirming theories.

When developing a test, the crucial question is the degree to which the test is a valid measure of the construct that we wish to assess. A construct is a latent or unobservable variable defined as an "attribute of people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283). The construct of interest in this case is the current behavioral and emotional adjustment of a selected child. Results obtained from the screener, therefore, would inform teachers, school officials, psychologists, and others (e.g., doctors, parents) about a student's behavioral and emotional (or in medical terms "health") status and guide decision-making and intervention accordingly.

Two essential steps in determining the usefulness of an instrument include assessing predictive validity, whether the scores from the screener predict important outcomes of interest, as well as assessing whether the screener can be used to differentiate between groups of children. By assessing these relationships, we are able to build and expand upon the nomological net of our proposed construct (Cronbach & Meehl, 1955) thus continuing to accumulate evidence to support the construct validity of a measure.

In order to assess known-groups validity, researchers often use an epidemiological screening model (Derogatis & DellaPietra, 1994). In this model the goal is to maximize the number of true positives and true negatives while minimizing false positives and false negatives. The hit rate is an overall measure of the proportion of cases correctly classified, including both true positives and true negatives. Sensitivity (true

positives) indicates the proportion of those individuals with emotional and behavioral problems who are detected by the screener. Specificity (true negatives) indicates the proportion of individuals without emotional and behavioral problems who are identified as such by the screener. When individuals without problems are identified by the screener as having problems, this misclassification is referred to as the false positive rate. These types of errors result in wasted resources and misidentification of children. False negatives occur when individuals who are having problems are not identified by the screener, leading to the denial of services to children in need. In screening, false positives are more acceptable than false negatives because it is preferable to identify individuals as needing further assessment when they actually do not, rather than allow individuals to suffer the consequences of mental illnesses with known morbidity without receiving treatment.

One can estimate the predictive power of a screener using the Positive Predictive Value (PPV) and Negative Predictive Value (NPV). PPV indicates the proportion of patients with positive screens who actually have emotional and behavioral problems. When the PPV is low a large number of false positives are present. On the other hand, when the PPV is optimized false positives are minimized at the risk of missing true cases. NPV indicates the proportion of patients with negative screens who actually do not have emotional and behavioral problems. When the NPV is low, a large number of false negatives result.

	Diagnosed	Not Diagnosed	
Positive Screen	True positive (a)	False positive (b)	Positive predictive value (PPV) a/a+b
Negative Screen	False negative (c)	True negative (d)	Negative predictive value (NPV) d/c+d
	Sensitivity a/a+c	Specificity d/b+d	Overall Hit Rate a+d/a+d+c+b

One must also keep in mind that the base rate of the outcome of interest will significantly affect the PPV and NPV of a screener (Meehl, & Rosen, 1955). As Hill, Lochman, Coie, Greenberg and the Conduct Problems Prevention Research Group (2004) so eloquently explained, "Sensitivity and specificity of tests may sound impressive when reported without reference to PPV, NPV, and base rates. For example, a test with sensitivity of .80 and specificity of .95 has a PPV of about 74% if the base rate is 15%, but the PPV is reduced to 46% if the base rate is 5%" (p. 810). A suggested estimate for an annual base rate of emotional and behavioral problems in a normative elementary school population from high-risk environments would be around 20% as supported by research (Hill, Lochman, Coie, Greenberg, & the Conduct Problems Prevention Research Group, 2004; Campaign for Mental Health Reform, 2005; Friedman, Katz-Leavy, Manderscheid, & Sondheimer, 1996); however, this base rate will be lower when focusing on a single disorder. Many screening research studies fail to provide PPVs and NPVs, limiting their reporting of findings to sensitivity and specificity. Thus very little guidance exists as to what constitutes adequate ROC curve analyses index values for a screening measure of emotional and behavioral adjustment.

Bennett and Offord (2001) have suggested that screening methods should have a minimal PPV and sensitivity of 50%, meaning at least 50% of the children labeled as high-risk are correctly classified (PPV) and at least half of the children with problems should be detected (sensitivity) in order to justify the use of the screener. Power and colleagues (1998) considered a cut off score clinically useful if PPV or NPV was great than or equal to .65 and if sensitivity or specificity was approximately .50 or greater. Other studies (Carran & Scott, 1992; Campbell, Bell, & Keith, 2001; Weis, Lovejoy, & Lundahl, 2005), on the other hand, indicated that sensitivity, specificity, PPV, and overall hit rate values should be equal to or greater than .80 to support the utility of a screening measure.

The usefulness of the screener for identifying children at risk for behavioral, emotional, and academic adjustment can be assessed using this conceptualization by performing a Receiver Operating Characteristic (ROC) curve analysis, comparing children with known problems versus those without. The ROC curve is a plot of the true positive rate against the false positive rate of different possible cut scores for a diagnostic test (Altman, 1991). ROC curves demonstrate the tradeoff between sensitivity and specificity (increases in sensitivity are accompanied by decreases in specificity) and use the area under the plotted curve as a measure of test accuracy. Results from a ROC curve analysis can be used to select an optimal cut score for identifying students at risk for developing emotional and behavioral problems.

The effectiveness of a screening measure is assessed by evaluating the accuracy of discrimination between children with emotional, behavioral, or academic problems and those without. An area under the curve (AUC) of 1 defines a perfect test, while an area

of .5 represents a relatively inefficient measure; ROC curve areas of .80-.90 are considered “good” discriminators while .90- 1 are considered “excellent.”

A Multi-disorder Screening Instrument

The existence of brief, multi-disorder screeners may provide an important piece of the technological infrastructure needed to convince school districts and health care providers that early identification is not only beneficial to children, but also practically delivered in schools and primary care settings. Traditionally, the content of emotional and behavioral screeners has been comprised of symptoms of disorders. When using symptom-based assessment to screen for a number of disorders, researchers often must sacrifice brevity and cost effectiveness in order to have broad coverage of symptomatology. Therefore, many symptom-based screeners focus on an individual disorder in order to maximize symptom coverage of that particular disorder. Although screening for symptoms of specific disorders indicates an important step in the acceptance of emotional and behavioral screening in general, this procedure also leads to a failure to identify large numbers of children who may have problems other than the target screening condition.

Theoretically, a multi-disorder screener is feasible if one invokes modern temperament and neurological theory and their variants (Gray, 1987; Rothbart & Bates, 1998). Although beyond the scope of this review there is an emerging consensus that much of the range of psychopathology seen in childhood is a function of the interplay of flawed emotional, behavioral, and attentional control systems. Further support for this point of view is the finding that co-morbidity is the rule in child psychopathology (Rutter & Sroufe, 2000). Yet further support can be found in the numerous factor analytic

studies of child behavior rating scales that produce three or four factor solutions (Reynolds & Kamphaus, 2004). These theoretical stances and associated factor analytic findings suggest that a screener that adequately assesses emotional, behavioral, and attentional control systems will be predictive of the onset of a variety of forms of psychopathology and other important outcomes.

For example, Leon, Kathol, Portera, Farber, Olfson, Lowell, & Sheehan (1999) conducted a large scale study of depression screening in a primary care setting. They found that a large number of patients with false positives met diagnostic criteria for other mental disorders thus indicating the need to take comorbidity into account and screen for general maladjustment rather than one or a limited number of disorders. Although the screener was meant to identify those with depression, it succeeded in identifying patients with other disorders as well due to overlapping symptomatology. As the first step in a multiple-gated system, screeners should simply identify those children with elevated symptomatology, leaving diagnosis of specific disorders to the later gates. Therefore, a need exists for the development of brief, multi-disorder child screening measures of emotional and behavioral adjustment.

Kamphaus, Thorpe, Winsor, Kroncke, Dowdy, and VanDeventer (2007) created one such measure: an abbreviated, 23-item version of the Behavior Assessment System for Children Teacher Rating Scale –Child Version (BASC TRS-C; Reynolds & Kamphaus, 1992). A predictive validity study was conducted using this screener with a two-year longitudinal sample of 206 children. Results indicated strong initial reliability and validity evidence. The internal consistency coefficient for screener scores was high at .97. The screener also predicted a substantial range of outcomes one year later (see Table

1) including conduct problems, social skills problems, depression, and achievement scores. Additionally, it predicted outcomes as well as an overall composite score based on a larger set of items that included both internalizing and externalizing items from the full BASC. These findings suggest that a multi- disorder screener can be both brief and predictive of a broad range of behavioral, emotional, and academic outcomes of importance.

Further evidence comes from a large scale study done by Goodman, Ford, Simmons, Gatward, and Meltzer in 2003. They found that the Strengths and Difficulties Questionnaire (SDQ), although meant to identify specific disorders, was much better at detecting children with more generalized symptomatology due to the high level of comorbidity as well as the overlap of symptomatology in child psychopathology. For example, children with developmental disorders were identified due to elevated emotional and hyperactivity symptomatology even though the SDQ does not contain 'core" autism spectrum symptoms.

Available Child Screeners

The following review of child screening instruments (see Table 2) is meant to be as comprehensive as possible; however, we do not suggest that the review is actually comprehensive as new instruments are developed every day. We focused on those instruments specifically developed for elementary school-aged children that had been the subject of research studies and contain information on psychometric properties in their manuals.

Child screeners have more variability than adult screeners due to the use of multiple informants (parent, teacher, clinician, self) as well as settings in which the

screener may be utilized (school, primary care). Multi-disorder screening measures of child behavior and emotional adjustment are rare, and those that do exist are often too long and time-intensive (more than 40 items) to be considered true screeners. Examples would include: the Achenbach Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1987), Behavior Assessment System for Children – 2 (BASC-2; Reynolds & Kamphaus, 2004), McDermott Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1994), Child/Adolescent Psychiatry Screen (CAPS), Swanson, Nolan, and Pelham Rating Scale -Revised (SNAP-IV-R; Swanson & Carlson, 1994), and the McCarney Behavior Evaluation Scale – 2 (McCarney & Leigh, 1990).

One measure that may be considered a true, multi-disorder screener is the Pediatric Symptom Checklist (PSC; Jellinek, Murphy, & Burns, 1986), a parent-report, 35 item symptom list developed from the lengthier Washington Symptom Checklist and used in primary care settings with school-aged children (ages 6-12). This measure has been extensively studied with a range of economically, racially, and clinically diverse samples and found to have strong internal consistency, test-retest reliability, interrater agreement, and validity for identifying children who would benefit from further, more intensive assessment (Jellinek, Murphy, & Burns, 1986; Jellinek, Little, Murphy, & Pagano, 1995; Jellinek & Murphy, 1988; Murphy, Reede, Jellinek, & Bishop, 1992; Simonian & Tarnowski, 2001; Walker, LaGrone, & Atkinson, 1989; Stoppelbein, Greening, Jordan, Elkin, Moll, & Pullen, 2005). It has been found to have good sensitivity, ranging from .77 to .95, and specificity, ranging from .68 to 1.0 (Stoppelbein, Greening, Jordan, Elkin, Moll, & Pullen, 2005; Jellinek, Little, Murphy, & Pagano, 1995; Walker, LaGrone, & Atkinson, 1989; Jellinek & Murphy, 1990; Simonian & Tarnowski,

2001;). Although designed for use in primary care settings, the PSC has also been shown to correlate highly with teacher ratings of child symptomatology and academic failure and has identified students whose difficulties were unknown to school staff thus suggesting that it may be of use in school settings as well (Murphy, Jellinek, & Milinsky, 1989); however, a teacher version of this instrument does not currently exist.

Two variations of the PSC have also been developed. Pagano, Cassidy, Little, Murphy, and Jellinek (2000) adapted the PSC into self-report format (Youth-PSC) and found that this measure correlated highly with teacher and parent ratings of child dysfunction as well as self-reported measures of depression and anxiety. It also demonstrated acceptable levels of sensitivity, 94%, and specificity, 88%, in identifying children at psychosocial risk (Pagano et al., 2000); PPV and NPV were not reported, however, the AUC was .66 which is lower than the .8 needed to be considered good. The PSC-Y identified children with internalizing symptoms that were missed by parents thus supporting the superiority of self-report measures in assessing internalizing symptoms. Gall, Pagano, Desmond, Perrin, and Murphy (2000) found support for the use of the PSC-Y in a high school-based health center environment as well.

Gardner, Murphy, Childs, Kelleher, Pagano, Jellinek, McInerny, Wasserman, Nutting, and Chiapetta (1999) created a short form of this instrument, PSC-17 which has demonstrated lower preliminary reliability estimates at .67 for the total score (Borowsky, Mozayeny, & Ireland, 2003). This instrument has been found to have adequate sensitivity at 82% and specificity at 81%; however, its PPV was found to be quite low at 15% (Gardner et al., 1999). Therefore, the authors warn that a positive screen “is not a

diagnosis," but rather a "signal for further examination of the child and family" as should be the case with all screening instruments (Gardner et al., 1999, p. 231).

The Strengths and Difficulties Questionnaire (SDQ) is a five-minute behavioral questionnaire containing 25 items that generate scores for Conduct Problems, Inattention-Hyperactivity, Emotional Symptoms, Peer Problems, and Prosocial Behavior as well as a Total Difficulties Score. This screener can be completed by parents or teachers of 4 to 16-year olds and also includes a self-report version for 11- to 16- year olds. The SDQ was developed in Great Britain based on theory using DSM-IV (APA, 1994) criteria as well as factor analyses. Since its development, the SDQ has been translated into 60 languages and extensively researched worldwide including Great Britain, Australia, Holland, Sweden, Norway, Germany, and Urdu (Goodman, 2001; Van Widenfelt, Goedhart, Treffers, & Goodman, 2003; Malmberg, Rydell, & Smedje, 2003; Ronning, Handegaard, Sourander, & Morch, 2004; Flawes & Dadds, 2004; Becker, Woerner, Hasselhorn, Banaschewski, & Rothenberger, 2004; Vostanis, 2006).

In several countries, the total score has been found to have adequate reliability with an alpha of .76 and test-retest of .96; however, the internal consistency of the individual scales, with the exception of the inattention-hyperactivity scale, have been questionable, especially Peer Problems with an alpha of .51 (Goodman & Scott, 1999; Mellor, 2004). In a British sample, Goodman and Scott (1999) found that the SDQ was significantly better than the CBCL at detecting inattention and hyperactivity, and as good at detecting both internalizing and externalizing problems indicating convergent validity.

In 2003, Goodman and colleagues performed a ROC curve analysis on a British community sample of 7984 5-15 year olds using the SDQ and found a sensitivity of

63.3%, specificity of 94.6%, PPV of 52.7%, and NPV of 96.4%. Sensitivity varied by diagnosis with 70-90% of conduct, hyperactivity, depression, developmental disorders, and some anxiety disorders being identified, but only 30-50% of those children with specific phobias, panic and eating disorders, and separation anxiety being identified.

This research seems to suggest that the SDQ would be best used as an indicator of general maladjustment with a second-gate being used to detect specific disorders. Additionally, one must also keep in mind that sensitivity is of the utmost importance when initially screening children for emotional and behavioral problems in order to minimize false negatives. False negatives should be minimal for a first gate screening instrument because we want to catch as many children with emotional and behavioral problems as possible at this stage. Children with emotional and behavioral problems who are missed at the first gate are not recoverable through later assessment.

An American version of the SDQ has been developed just recently and preliminary findings are positive (Bourdon, Goodman, Rae, Simpson, & Koretz, 2005). As opposed to the five factor structure found in England, Dickey and Blumberg (2004) found a stable three factor model in a US sample consisting of internalizing problems, externalizing problems, and a positive construal factor consisting of prosocial items. Two NIMH – funded studies are currently under way to examine the validity of the SDQ. The worldwide interest in the SDQ and extensive research currently being done provides an excellent opportunity for researchers to examine cross-cultural similarities and differences with regard to psychosocial adjustment.

Several child emotional and behavioral screeners consist of a number of quick screens for multiple disorders. For example, the Beck Youth Inventories of Emotional

and Social Impairment (BDI; Beck, Beck, & Jolly, 2001) are designed for children ages 7 through 14 years and consist of five 20-item self-report scales that assess symptoms of depression, anxiety, anger, disruptive behavior, and self-concept. These scales can be used separately or in combination depending on the child's individual needs and time constraints. Bose-Deakins and Floyd (2004) found these scales to have adequate reliability and convergent validity; however, they noted several cautions and limitations regarding the available validity evidence. First, a principal axis factor analysis suggests that the majority of the inventories, including Anxiety, Depression, and Anger, appear to measure the same general construct of negative affect. Additionally, evidence of the inventories' abilities to discriminate between children with emotional and behavioral problems and those without as well as discriminating between different emotional and behavioral problems is lacking. Thus, these inventories are best used as a first gate in order to assess general risk for emotional and behavioral problems. More research on the validity of these scales should also be done as only one published study was found (Steer, Kumar, Beck, & Beck, 2005), providing construct validity evidence consistent with that found in the manual.

The DISC (Diagnostic Interview Schedule for Children) Predictive Scales – version 4.32 (DPS-4.32; Leung, Lucas, Hung, Kwong, Tang, Lee, Ho, Lie-Mak, & Shaffer, 2005) was recently updated to include work done on the NIMH DISC-IV (Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000), reflecting DSM-IV diagnostic criteria. The DPS – 4.32 consists of parent (14 scales with total of 92 items) and youth (18 scales with total of 98 items) questionnaires that assess the likelihood of a young person, ages 8 to 18, having any of 18 disorders. Additionally, the DPS provides a

separate impairment module indicating the degree to which a behavior is having a negative impact on the individual's social, academic and family life. The items were derived from the full DISC (Schwab-Stone, Shaffer, Dulcan, Jensen, Fisher, Bird, Goodman, Lahey, Lichtman, Canino, Rubio-Stipec, & Rae, 1996), by identifying those items that were most predictive of specific diagnoses (Lucas, Zhang, Fisher, Shaffer, Regier, Narrow, Bourdon, Dulcan, Canino, Rubio-Stipec, Lahey, & Friman, 2001).

In the original version (DPS-2.3), the substantial reduction in scale length was not associated with any significant changes in discriminatory power. Lucas and colleagues (2001) examined the DPS-2.3 classification accuracy for a number of disorders including simple phobia, social phobia, agoraphobia, OCD, MDD, ADHD, ODD, and conduct disorder. They found adequate reliabilities, sensitivities ranging from .67 to 1.00, specificities from .49 to .96, PPV from .07 to .74, and NPV from .87 to 1.00. They concluded that the DPS is a valuable tool for determining subjects who do not need further assessment and speeding up the structured diagnostic interviewing process; however, external validity studies were lacking.

An examination of the psychometric properties of the new parent DPS- 4.32 version using a community sample (N=541) of Chinese children found adequate reliability as well as adequate specificity (.91), and NPVs (.98); however, sensitivity was a bit low at .68 and PPV was found to be .34. Once again, more research should be done to reinforce these findings on other samples (Leung, Lucas, Hung, Kwong, Tang, Lee, Ho, Lie-Mak, & Shaffer, 2005).

Other child emotional and behavioral screeners tend to focus on one or several specific diagnoses or problems. The Yale Children's Inventory (YCI; Shaywitz, Schnell,

Shaywitz, & Towle, 1986) is a parent-rated scale based on the *Diagnostic and Statistical Manual of Mental Disorders 3rd edition* (DSM-III; APA, 1987) that focuses on the assessment of learning disabilities, emphasizing attention deficits. Scale development was based on factor analyzing items that differentiated children with learning disabilities from a control group in a community setting, especially focusing on inattention items (Shaywitz et al., 1986).

In assessing the ability of the YCI's attention scale to discriminate between ADHD and normals, the scale was found to have a sensitivity of 87.5% and specificity of 94% (Olafsen & Sommerfelt, 1999). Shaywitz, Shaywitz, Schnell, and Towle (1988) found evidence of both concurrent and predictive validity with the YCI correlating significantly with both cognitive and behavioral outcomes. The YCI has also been researched extensively cross-culturally (Olafsen & Sommerfelt, 1999). In reviewing the literature, Olafsen and Sommerfelt (1999) concluded that the YCI may be a valuable tool for the early identification and screening of Norwegian children with attention deficit disorders. The authors may want to consider revising the YCI in order to reflect DSM-IV diagnostic criteria.

Externalizing disorders, especially Attention Deficit Hyperactivity Disorder (ADHD), have been the focus of numerous screening measures for children. The Conners Rating Scales – Revised (CRS-R; Conners, 1973; Conners, Parker, Sitarenios, & Epstein, 1997) are symptom- based rating scales that are widely- used in schools, mental health clinics, residential treatment centers, pediatric offices, juvenile detention facilities, child protective agencies, and outpatient settings to screen for ADHD, learning problems, and conduct problems. The authors have suggested that the CRS-R may be used as a

screening measure as well as a tool for treatment monitoring, a diagnostic aid, and a research instrument. There are three versions—parent, teacher and adolescent (ages 12 through 17) self-report—all of which also have short (parent: 27 items; teacher: 28 items; self: 27 items) and long (parent: 80 items; teacher: 59 items; self: 87 items) forms available. The long forms are too extensive to be used as screening measures; however, users also have the option of administering a 12-item ADHD Index or the 18-item DSM-IV Symptom Checklist, or both. This instrument has been found to have adequate reliability and validity (Conners et al., 1997), but has been criticized for having too low cut-off scores thus inflating prevalence rates. However, classification indices are quite high with sensitivities ranging from .78 to .92, specificities ranging from .84 to .94, PPV ranging from .83 to .94, and NPV ranging from .81 to .92 depending on informant (parent, teacher, adolescent) (Conners et al., 1997).

The AD/HD Comprehensive Teacher's Rating Scale (ACTeRS**;** Ullman, Sleator, & Sprague, 1988**)** is a 24-item teacher-rated ADHD screener for children from kindergarten through 5th grade. Although this scale has adequate reliability, it has not been widely researched and contains little supportive data in the manual concerning validity. The manual also lacks information regarding the standardization sample. Ullman, Sleator, and Sprague (2000) found that the ACTeRS could differentiate between children with and without ADHD as well as children with Learning Disabilities and those with ADHD; however, convergent and divergent validity has not been examined. Additionally, the ACTeRS does not reflect current subtypes of ADHD as discussed in the DSM-IV (APA, 1994; Demaray, Elting, & Schaefer, 2003). Therefore, this instrument is not recommended for diagnosing ADHD. Although it has not been validated as a

screening measure, the ACTeRS may serve this purpose more effectively since it has been found to discriminate between children with and without ADHD. Research should be done to examine this possibility.

The ADHD Rating Scale – IV (ADHD-IV; DuPaul, Power, Anastopoulos, & Reid, 1998) is an 18 item rating scale for children ages 5 to 18, containing both parent and teacher versions. It is based upon DSM-IV diagnostic criteria and contains Inattention and Hyperactivity subscales. The ADHD-IV was standardized on a large nationally representative sample, and the manual provides excellent reliability and validity (content, internal structure, convergent and divergent, predictive) evidence (DuPaul et al., 1998). The manual also provides different cut-off scores depending on the purpose of the assessment (rule-out/screening vs. diagnosis). Parent ratings have sensitivities of .83 to .84, specificities of .49, PPV of .54 to .58, and NPV of .77 to .81. Teacher ratings produce sensitivities of .63 to .72, specificities of .86, PPV of .78-.79, and NPV of .73 to .81 (DuPaul et al., 1998). In general, the ADHD-IV is a well-developed instrument that could be used to screen school aged children for ADHD; however, Collett, Jeneva, and Myers (2003) warn users about the risk of misclassifying youth due to suboptimal sensitivity and specificity.

The Eyberg Child Behavior Inventory (ECBI; Eyberg & Pincus, 1999) is a parent-rated 36 item questionnaire designed for use in pediatric settings as a quick screen for disruptive behavior in children ages 2-16. The Sutter-Eyberg Student Behavior Inventory – Revised (SESBI-R; Eyberg & Pincus, 1999) was created during the 1999 revision of the ECBI as a teacher rated version and consists of 38 items, 13 of which are new to the SESBI, replacing non-school related items from the ECBI. The standardization of the

SESBI-R is problematic, consisting of 415 elementary school children from 11 schools in Gainesville, FL (Meikamp, 2003). The SESBI-R is supposed to target children ages 2-16, however not all ages were represented in the norming samples.

The ECBI has been found to have adequate reliability and concurrent validity. (Boggs, Eyberg, & Reynolds, 1990). The ECBI was also found to discriminate between normal and conduct-problem adolescents (Eyberg & Robinson, 1983). Rich and Eyberg (2001) found the ECBI to have excellent classification accuracy in a sample of children ages 3 to 6 with a sensitivity of .96, specificity of .87, PPV of .88, indicating that 88% of the children who exceeded the cut-off score were correctly identified, and NPV of .96.

Weis, Lovejoy, and Lundahl (2005) found the ECBI to be useful for screening children for externalizing disorders, but less useful in discriminating between specific behavior problems. When classifying children with specific externalizing behavior problems, sensitivities ranged from .63 for the Conduct problem component of the ECBI to .77 for the Inattentive component. Specificities were all above .90. They found that all components of the ECBI displayed adequate NPV, ranging from .82 to .94. The ECBI Inattentive and Oppositional components displayed PPV of .85 and .80 respectively, while the Conduct problem component exhibited lower PPV at .63.

The SESBI-R has some preliminary reliability and validity evidence; however, no reliability or validity evidence exists for older children (Whiston & Bouwkamp, 2003). More research is needed on the SESBI-R.

Other measures focus on internalizing symptoms such as anxiety and depression. These include self-report measures for school-aged children and adolescents such as the Reynolds and Richmond Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds

& Richmond, 1985), the State-Trait Anxiety Inventory for Children (STAIC; Spielberger, 1973), the Multidimensional Anxiety Scale for Children (March, Parker, Sullivan, Stallings, & Conner, 1997), the Reynolds Child Depression Scale (RCDS; Reynolds, 1989), and the Children's Depression Inventory (CDI; Kovacs, 1992).

The STAIC and RCMAS have been found to have good reliability and criterion-related validity. These tests can differentiate between youth with anxiety disorders and no disorder; however, findings are mixed on their ability to discriminate between diagnostic groups, especially between internalizing problems such as anxiety and depression. (Kamphaus & Frick, 2002; Seligman, Ollendick, Langley, & Baldacci, 2004). This may be due to item content and overlap with depression measures such as the CDI. Seligman and Ollendick (1998) found that approximately 21% of RCMAS items and 25% of STAIC items overlapped with items on the CDI. Thus the STAIC and RCMAS may be best used as first gate screeners in a multiple-gate system even though they were not developed and validated for this purpose. More research is needed to examine the utility of these instruments in a screening capacity.

The Multidimensional Anxiety Scale for Children (March, Parker, Sullivan, Stallings, & Conner, 1997) is a newer self-report anxiety measure for youth ages 8 to 19 consisting of 39 items. A 10 item short form of this measure also exists; however, this form has been found to have low reliability and lacking of validity evidence (Caruso, 2001). The MASC has been found to have adequate reliability, including test-retest reliability (March & Sullivan, 1999; Christopher, 2001), as well as good convergent and divergent validity (March, et al., 1997). Rynn, Barber, Khalid-Khan, Siqueland, Dembiski, McCarthy, and Gallop (2006) used the MASC to discriminate between

children with Generalized Anxiety Disorder and Depression. They found the AUC of .623 to be in the poor to fair range. When sensitivity was set at .80, maximum specificity was found to be .34. This instrument has not been validated as a screening instrument in a multiple-gate screening system.

The CDI is a 27-item self-report depression measure for youth ages 8 to 17 modeled after the adult Beck Depression Inventory. Parent and teacher forms also exist, but little research has been done on them. Unfortunately, reliability and validity findings for the CDI are mixed. Several studies have found that the CDI does not possess the properties of a useful screening tool when using suggested cut-offs of 13 with sensitivity at .40, specificity at .729, PPV at .136, and NPV at .92 or 20 with sensitivity at .142, specificity at .906, PPV at .138, and NPV at .909 (Matthey & Petrovski, 2002; Kresanov, Tuominen, Piha, & Almqvist, 1998). Timbremont, Braet, and Dreessen (2004), on the other hand, concluded that the CDI did have adequate psychometric properties to be used as a screening tool to select potential cases for further assessment for depression with a specificity of .84, sensitivity of .94, PPV of .63 and NPV of .98. Until more support is found for the psychometric properties of the CDI, this instrument should be used with great caution (Kavan, 1992; Knoff, 1992).

The RCDS (Reynolds, 1989) is another self-report measure intended to assess the severity of depressive symptomatology in children ages 8 to 12. Sensitivity of .73 and specificity of .97 are reported (Reynolds, 1989). This measure has strong reliability and validity evidence with the exception of discriminant validity as it correlates highly with anxiety measures (Kamphaus & Frick, 2002), and is advertised for use as a large scale screening instrument.

The Center for Epidemiological Studies Depression Scale Modified for Children (CES-DC; Faulstich, Carey, Ruggiero, Enyart, & Gresham, 1986) was adapted from the adult CES-D. This scale lacks reliability and validity evidence, and has not been adequately researched. Faulstich and colleagues (1986) found that the measure had poor reliability and validity for children, thus requiring more validational support before it can be recommended for use. Another scale, the Columbia Depression Scale (CDS) is a 22-item self-report scale, derived from the Major Depression section of the Diagnostic Interview Schedule for Children (Shaffer, Fisher, Lucas. Dulcan. & Schwab-Stone, 2000); however, this scale is lacking reliability and validity evidence.

The most severe outcome of mental illness is suicide. As mentioned earlier, suicide has emerged as the third leading cause of death in youth ages 15 to 24. Furthermore, over 90% of children and adolescents who commit suicide have at least one mental disorder, the most common type being mood disorders (Campaign for Mental Health Reform, 2005; Shaffer, Scott, Wilcox, Maslow, Hicks, Lucas, Garfinkel, & Greenwald, 2004). As Shaffer and colleagues (2004; p. 71) reasoned, "If the risk factors for suicide are both identifiable and treatable, screening teens for untreated mood disorders should be an important component of any suicide prevention program."

A number of screening instruments have been developed in order to assess suicidal risk in adolescents including the Beck Scale for Suicidal Ideation (Beck, Kovacs, & Weissman, 1979), the Suicide Risk Screen (Eggert, Thompson, & Hering, 1994), and the Suicidal Ideation Questionnaire (Reynolds, 1988, 1991) which yielded adequate sensitivity ranging from 83% to 100% with less than adequate specificity from 40% to 70% in a Midwestern US high school. The Beck Scale for Suicidal Ideation provides no

reliability and validity information for adolescents, and therefore should not be used until this information is collected. The Suicide Risk Screen assesses suicide ideation, suicide attempts, depression, and substance use, all factors found to predict suicide (Shaffer et al., 2004; Brent, Baugher, Bridge, Chen, & Chiappetta, 1999) in adolescents 14 years and older. Thompson and Eggert (1999) found the Suicide Risk Screen to have sensitivity ranging from 87% to 100%, but low specificity from 54% to 64% in a sample of 581 high school youth.

The Columbia Suicide Screen (CSS; Shaffer et al., 2004) is a 14- item self-report questionnaire that assesses the most important risk factors for suicide in youth ages 11 to 18. These items are embedded within a larger screen of general health and relationship items, the Columbia Health Screen, in order to avoid a focus on suicide. Shaffer and colleagues (2004) found this instrument to have adequate sensitivity, .75, in identifying high schoolers at-risk for suicide; however, they did recommend a second stage of evaluation in order to "reduce the burden of low specificity" even though the specificity of .83 is superior to most other instruments (p. 71). The PPV was very low at 16% which would result in 84 false positives for every 16 youths correctly identified. This amount and type of misidentification may not be acceptable to schools and parents. In general, most suicide screens are limited to adolescent and adult populations and suffer from low specificity which may overburden programs with false positives. Thus these screens should only be used as first gates in a multi-gate system.

As stated earlier, this review of the available screening instruments is far from exhaustive; additionally, a word of caution is in order. Although an exorbitant number of instruments exist, one must be careful to assess each instrument's psychometric

properties before choosing to utilize that instrument. Many of the instruments reviewed above, as well as those left unmentioned, still need more research done before one can be truly confident in their psychometric properties as screeners for emotional and behavioral adjustment. Additionally, one must remember that these instruments are NOT diagnostic, but rather should be used as indicators for further assessment.

Implementation of a Universal Screening System

Multiple gate screening procedures

Multiple- gated identification procedures are often used when implementing a universal screening program and are generally aligned with acceptable principles of prevention science (Weisz, Sandler, Durlak, & Anton, 2005). A multiple-gated identification procedure begins by screening an entire population for emotional and behavioral difficulties (universal screening). Those students identified by the screening instrument as being at-risk for emotional and behavioral problems are then assessed again using a different, often more thorough, assessment tool such as a full behavior rating scale (selected assessment). Lastly, the students who are identified by the second assessment as having emotional or behavioral problems receive a more comprehensive, individual assessment (indicated assessment). In this way, multiple-gating narrows down the population sequentially so as to yield groups of successively more impaired students. Those children identified as “at-risk” by the screener may benefit from selective prevention strategies while those identified as having more severe impairments would be referred for more intensive interventions. This type of procedure should increase identification and diagnostic accuracy as well as reduce costs due to inefficient identification (Hill, Lochman, Coie, Greenberg & The Conduct Problems Prevention

Research Group, 2004; Walker & Severson, 1990; Lochman & The Conduct Problems Prevention Research Group, 1995).

The Systematic Screening for Behavior Disorders (SSBD; Walker & Severson, 1990) is a multiple-gated procedure attempting to identify students in elementary school who are at elevated risk for externalizing or internalizing behavior disorders. The SSBD is one of the only multiple-gated procedures found in the literature that is designed to screen for multiple adjustment problems in children as opposed to a single disorder such as ADHD. This screening procedure consists of three stages: 1. Teacher ranking of all students in the classroom according to the externalizing and internalizing dimensions, 2. Teacher completion of behavior rating scales for the top three "internalizers" and "externalizers" in the classroom, and 3. Direct observation of those students above the Stage 2 cutoff score using a classroom and playground observational code. Researchers have found the SSBD procedure to be valid and reliable, as well as cost-efficient, in identifying children in need of services; additionally, the SSBD has been rated favorably by study participants including teachers and psychologists (Philips, Nelson, & McLaughlin, 1993; Walker & Severson, 1994). However, multiple gate systems that include teacher training and rankings (i.e., nomination) and classroom observations are relatively expensive in terms of personnel costs as assessed by teacher and other staff time devoted to this task. In addition to the time spent actually completing the task, teachers and observers must be also trained.

August, Ostrander, and Bloomquist (1992) utilized a multiple-gated identification system in order to assess for attention-deficit hyperactivity disorder in 1,490 elementary school students. In their procedure, Stage 1 consisted of teachers' ratings of the child's

behavior using the Child Behavior Checklist Teacher Report Form (CBCL – TRF) (Achenbach, 1991), Stage 2 consisted of parents' ratings using the parent version of the Child Behavior Checklist (Achenbach, 1991), and Stage 3 involved the administration of a structured psychiatric interview, the Diagnostic Interview for Children and Adolescents Revised – Parent version (DICA-R-P; Reich & Welner, 1990). When the child obtained a T-score of 60 or greater on the Attention Problems scale on the CBCL - TRF, then a CBCL was administered to a parent or guardian. If a T-score of 65 or greater was obtained on the Attention Problems scale of the parent form, then the DICA-R-P was administered at stage three of the assessment. The procedure resulted in an excellent PPV with 90% of the children identified in Stage 2, subsequently receiving an ADHD diagnosis at Stage 3, thus suggesting that the three-stage screening procedure maximized the use of time necessary to diagnose ADHD (August et al., 1992). However, the length of the rating scales used in this study is of great concern, especially if this procedure was to be implemented on a large scale.

August, Realmuto, Crosby, and MacDonald (1995) also employed a multiple-gate screening procedure to identify children at risk for conduct disorder. Once again, they employed a three gate procedure; however, they chose to utilize a specific section of the Conners Rating Scale in this study. Gate 1 consisted of teachers completing the 10-item Hyperactivity Index of the Revised Conners Teacher Rating Scale for the entire population (CTRS-R; Goyette, Conners, & Ulrich, 1978). In Gate 2, parents of those children who received a score of 1.6 or higher completed the 10-item Hyperactivity Index of the Revised Conners Parent Rating Scale (CPRS-R; Goyette et al., 1978). Lastly, a set of 15 items was given to assess parent behavioral management practices. It was found

that the procedure adequately discriminated children with higher adjustment from those with lower adjustment with all measures contributing significantly to prediction of child's self-concept, problem behaviors, and social skills (August et al., 1995). Additionally, the procedure predicted diagnostic ratings of psychiatric symptomatology with Gate 1 predicting both ADHD and ODD while Gate 2 added to the prediction of ADHD, but not ODD. Gate 3 contributed to the prediction of ODD, but not ADHD. Thus, it appears that the addition of gates and informants aided in accurate identification of children with behavioral and emotional maladjustment, depending on the diagnosis of interest.

Limitations of these two studies include:

1. Focused on externalizing symptomatology
2. Did not use actual screening measures, but rather selected items from longer measure based on content or an entire longer behavior rating scale
3. Did not look at the utility of adding a second, more comprehensive measure as a second gate, rather simply used a different informant as second gate
4. Did not look beyond a Gate 1: Teacher Screen, Gate 2: Parent Screen multiple gate procedure to examine different combinations of informants and gates. Thus one cannot assess whether the additional variance explained by the second gate is due to adding a second gate or second informant or both.

The Conduct Problems Prevention Research Group (Hill, Lochman, Coie, Greenberg, & The Conduct Problems Prevention Research Group 2004; Lochman and the Conduct Problems Prevention Research Group, 1995) has conducted a number of screening procedure studies, focusing on externalizing symptomatology; however, it must be noted that these studies did not utilize an actual multiple-gate procedure. Rather than

narrowing down the pool of individuals successively at each gate, these studies simply analyzed whether adding additional screening measures to a regression equation significantly increased screening accuracy on the full sample. They were actually answering the question of whether two measures are needed at the first screening point, not whether a second screening point is needed. Multiple regression analyses such as these can inform research on multiple gate procedures, but should not be considered as multiple-gate studies.

Hill, Lochman, Coie, Greenberg, and The Conduct Problems Prevention Research Group (2004) compared the effectiveness of single versus multiple raters and single versus multiple time-points when screening for externalizing problems. They did not utilize actual screening instruments, but rather selected externalizing items from the Teacher Observation of Classroom Adaptation – Revised (TOCA-R; Werthamer-Larsson, Kellam, & Wheeler, 1991) and the Child Behavior Checklist and Revised Problem Behavior Checklist (Achenbach, 1991). This study compared the predictive validity of six different screening models using logistic regression and ROC curve analyses:

1. Teacher Kindergarten (K)
2. Teacher 1st grade
3. Teacher K + Teacher 1st
4. Teacher K + Parent K
5. Teacher K + Parent 1st
6. Teacher K + Teacher 1st + Parent K + Parent 1st

They found that single time-point, multiple rater (parent and teacher) screening (Model #4) was the most effective and efficient in predicting externalizing outcomes; however,

this study failed to examine the utility of having a parent screen as the first gate, and did not use a more comprehensive measure as the second gate, as they focused on different time-points rather than different instruments. In this way, they actually did not have a two-gate, same informant combination because the teacher who rated the child in first grade was a different teacher than the one who rated him or her in Kindergarten.

Lochman and the Conduct Problems Prevention Research Group (1995) also found that adding a parent screen to a first-gate teacher screen accounted for a significant amount of the variance beyond that explained by just the teacher screen. Like the above studies, Lochman and colleagues (1995) pulled items from longer measures in order to create the screening measures rather than assessing the utility of actual screening measures and did not use a more comprehensive measure as the second gate. Additionally, like August and colleagues (1995), this study did not go beyond the Gate 1: Teacher Screen, Gate 2: Parent Screen multiple gate procedure to examine different combinations of informants and gates.

As one can see, a number of variables must be considered when developing multiple gate screening procedures. One must evaluate the utility of multiple gates as well as what informants should constitute each gate. One could view multiple informant assessment as a special type of multiple gate assessment when different informants constitute the different gates or levels. These two variables often are confounded in research studies, making it difficult to examine the relative utility of each separately; however, these variables must be evaluated separately in order to discern what combination of gates and informants is most efficient and accurate for a given purpose. Most research studies have not systematically varied gates and informants in order to

examine the relative accuracy of specific gate/informant combinations; rather, they simply have focused on whether a certain multiple-gate procedure was valid in general without testing it against other versions of the procedure (i.e., different informants, number of gates). As called for by Johnston and Murray (2003), "future research needs to address …the value of different informants at various stages of the assessment process…" (p. 500).

Number of Gates

Currently, no consensus exists as to how many levels of assessment are optimal for identifying those children in need of mental health services. In contrast with many of the above studies focused on three gate procedures, Simonian and Tarnowski (2001) have suggested a two-stage multimethod system of initial identification of risk (i.e., brief, cost-efficient screening) and subsequent diagnostic assessment (i.e., more comprehensive, multimethod battery) of children in pediatric settings with mental health needs. Pagano and colleagues (2000) discussed the implementation of a similar two-stage model in educational settings.

As described above, Lochman and the Conduct Problems Prevention Research Group (1995) examined the usefulness of combining two screening measures, in screening kindergartners for 1st grade adjustment problems and found that the second gate of the screening procedure (parent ratings) clearly added to the effectiveness of the first gate (teacher screen) in predicting problem behavior; however, adding a third instrument that measured parent practices did not significantly aid in prediction. In this study, one cannot distinguish whether the increase in effective classification was due to the addition

of the second level, a different informant, or both. Additionally, one must keep in mind that this study did not use an actual multiple gate procedure.

In contrast to Lochman and the Conduct Problems Prevention Research Group's (1995) findings, a three gate system implemented by Loeber, Dishion, and Patterson (1984) consisting of a) teachers' ratings of children's problem behaviors, b) mothers' ratings of children's problem behaviors, and c) mother's reports of child-rearing strategies, was found to predict no better than a single screening gate using the Aggression, Moody, Learning Disability teacher rating scale (AML; Cowen, Dorr, Clarfield, Kreling, McWilliams, Pokracki, Pratt, Terrell, & Wilson, 1973; Carberry & Handal, 1980). Thus more research is needed to discern the optimal number of gates for a multiple-gate procedure.

Multiple informants

In child assessment and diagnosis, it is often recommended that ratings be collected from multiple informants including parents, teachers, as well as the youth themselves so as to provide the greatest amount of information possible from which to make decisions (Kamphaus & Frick, 2002). In a multiple-gated screening system, we have the opportunity to implement this recommendation across gates or levels of assessment; however, several issues exist when attempting to integrate and interpret ratings of multiple informants.

When numerous informants indicate a similar problem, then the clinician can feel more confident in the validity of his or her diagnosis; however, a lack of consistency across ratings of different informants is more often the case, with low agreement among informant ratings from different settings (e.g., parents and teachers, parents and children)

and modest agreement among informants from similar settings (e.g., mother and father, two teachers) (Youngstrom, Loeber, & Stouthamer-Loeber, 2000; Achenbach, McConaughy, & Howell, 1987; van der Ende, 1999; Grietens, Onghena, Prinzie, Gadeyne, Van Assche, Ghesquiere, & Helinckx, 2004; Grills & Ollendick, 2003). In their classic meta-analysis, Achenbach and colleagues (1987) found that the mean correlation between mothers and fathers was .59, between parents and teachers was .27, and between children and other informants was only .22. This low agreement indicates that informants are not interchangeable thus suggesting the need for multiple informant ratings; however, questions still exist regarding why these discrepancies exist as well as what might be the best way to integrate conflicting information.

Several possible explanations may be offered for the low rate of agreement between informants, none of which are mutually exclusive. In fact, it is more likely that a number of these explanations act in unison to produce informant disagreement. As Renk (2005) explains, "Such disagreements may be viewed as bias or error on the part of one of the informants, as support for the variability of children's behavior across situations, as an informant's lack of access to certain types of behavior, as denial of the behavior of interest, or as distortion of information by an informant."

First, one must consider the possibility that parents, teachers, and children each provide unique, meaningful information. Parents and teachers see the child in different settings, and their ratings may reflect true behavior variations across these settings. For example, discrepant inattention ratings between parents and teachers may reflect different demands at home and at school. Achenbach and colleagues (1987) provided evidence for

the possibility of actual situational variability by showing that informants in different settings show much lower correlations than those in similar settings.

Ratings also depend on informant characteristics. An informant's impression of another individual is based upon his or her interpretation of that individual's behavior. Thus all ratings are subject to the characteristics and judgments of the rater. As van der Ende (1999) pointed out, "each individual holds different thresholds and personal standards when rating problem behavior" which depend on their knowledge of what constitutes normal behavior, expectations of the child, as well as access to a same-aged peer group from which to compare. Teachers have the advantage of observing the child within a peer group, thus allowing them to distinguish between maladaptation and normal, age-related problem behavior. Other informant variables that affect ratings include personality characteristics, psychopathology such as depression (Boyle & Pickles, 1997; Youngstrom et al., 2000; Clarke-Stewart, Allhusen, McDowell, Thelen, & Call, 2003), as well as the informant's own motives, biases, and expectations (Renk, 2005; Grietens et al., 2004). Additionally, the parent-child relationship has been found to affect parent ratings of child behavior (van der Ende, 1999; Kamphaus & Frick, 2002; Clarke-Stewart et al., 2003). Thus one must distinguish whether informants are providing valid information before concluding that these discrepancies support the value of collecting information from multiple informants.

Studies have found that agreement between informants varies depending on several factors including the nature of the problem being assessed (Achenbach, McConaughy, & Howell, 1987; Grills, & Ollendick, 2003; Mesman, & Koot, 2000; Loeber, Green, Lahey, & Stouthamer-Loeber, 1991; Sourander, Helstela, & Helenius,

1999; Youngstrom, Loeber, & Stouthamer-Loeber, 2000), the clinical status of child (Handwerk, Larzelere, Soper, & Friman, 1999), the informants' psychological functioning (Youngstrom, Loeber, & Stouthamer-Loeber, 2000), and the age of the child (Achenbach, McConaughy, & Howell, 1987; van der Ende, 1999). In general, agreement tends to be higher for externalizing problems than for internalizing problems (Achenbach et al., 1987; Kolko & Kazdin, 1993; Grietens et al., 2004) and for younger children than for older children (Achenbach et al., 1987; van der Ende, 1999).

Clinicians have been found to weigh adult ratings, such as teachers and parents, more heavily for externalizing behaviors and child self-report more heavily for emotional or internalizing problems (Loeber, Green, & Lahey, 1990). Research has supported these decisions, finding that internalizing problems are best identified through self-report (Pagano et al., 2000; Youngstrom, Loeber, & Stouthamer-Loeber, 2000; Loeber, Green, Lahey, & Stouthamer-Loeber, 1991), and that children often report less externalizing problem behaviors than either parents or teachers. Correlations between child-reported internalizing syndromes and parent- and teacher- reported syndromes have been found to be low to medium at best (Kolko & Kazdin, 1993; Mesman & Koot, 2000).

Youth self-report of both externalizing and internalizing symptomatology may also become more valuable as the youth gets older as younger children may not have developed the abilities necessary to accurately reflect on and report feelings and behaviors (Grills & Ollendick, 2003). Sourander and colleagues (1999) found that a community sample of adolescents reported more internalizing and externalizing problems than did their parents, including suicidal symptomatology. However, studies involving clinical adolescent samples have found the opposite effect with parents rating the

adolescent's problems as more severe than the adolescents who tended to minimize problem behavior (Handwerk, Larzelere, Soper, & Friman, 1999). Thus the clinical status of the adolescent is an important variable to consider when deciding the informants on which to base diagnostic decisions.

Contradicting opinions exist as to the superiority of parent or teacher ratings. Several studies (Youngstrom et al., 2000; Loeber et al., 1991) found that teachers reported fewer internalizing and externalizing problems than did caregivers or youth; on the other hand, Kaufman, Cook, Arny, Jones, and Pittinsky (1994) found that teachers identified *more* problems. Several studies have found mothers to be more accurate in perceiving internalizing problems in children than teachers (Loeber et al., 1990; Youngstrom et al., 2000; Grietens, et al., 2004); however, Mesman and Koot (2000) found that teachers were more likely than parents to notice internalizing problems.

Wolraich, Lambert, Bickman, Simmons, Doffing, and Worley (2004) found that teachers reported higher levels of inattention than parents, perhaps due to differing environmental demands. Reynolds and Kamphaus' (1992) research with teacher rating scales has demonstrated that, on average, teacher ratings of child behavior are more reliable than parent ratings at preschool, child, and adolescent age levels, and that different teachers rate the same child similarly. Loeber and colleagues (1991) found teacher reports, as compared to child and parent reports, of ADHD symptoms in elementary school children to be the best predictors of later impairment including suspensions and special education placement. Goodman and colleagues (2004) found that teachers and caregivers provided information of roughly equivalent predictive value.

Additionally, contradictory opinions exist as to the utility of collecting ratings from multiple informants. Several studies (Lochman and the Conduct Problems Prevention Research Group, 1995; Biederman, Keenan, & Faroane, 1990) found that adding another informant added little variance to the identification process beyond that provided by the first informant. Jones and colleagues (2002) found that the effect of combining parent and teacher ratings was equal to or minimally higher than that of the teacher-only rating. These findings would indicate that multiple informants are not necessary in a multiple gate screening system; rather one could simply have the same informant (e.g. teacher) complete a more comprehensive rating scale for the second gate.

Goodman and colleagues (2003), on the other hand, found that the SDQ predicted best when ratings by all possible informants (parents, teachers, and child) were taken into account. Power and colleagues (1998) concluded that a combined informant approach, parent and teacher, was more successful in predicting the presence of ADHD than the single informant approach. Hill and colleagues (2004) found parent-teacher models to be superior; however, teacher-only models did have good predictive value for both externalizing and delinquency outcomes. Goodman and colleagues (2004) found that SDQ prediction was best when both caregiver and teacher ratings were completed. They also found that self-reports provided little extra information above that provided by either parent or teacher ratings. These findings would support the use of a different informant for the second gate of a multiple gate screening system.

In general, the majority of researchers and clinicians continue to emphasize the importance of multi-informant assessment (Verhulst, Dekker, & van der Ende, 1997; Power, Andrews, Eiraldi, Doherty, Ikeda, DuPaul, 1998; Jensen, Rubio-Stipec, Canino,

Bird, Dulcan, Schwab-Stone, & Lahey, 1999; Kamphaus & Frick, 2002); however, one must keep in mind that a consensus has not yet been reached on this issue.

It is often assumed that more is always better, whether it be number of informants, methods, or levels of assessment; however, this has not always been found to be the case (McFall, 2005). Adding the results of less reliable and valid measures does not increase predictive accuracy, but rather leads to contamination of findings. Furthermore, one reaches a point where adding more measures no longer contributes enough unique variance to be worth the effort. Specific combinations of measures, including different informants and number of gates, must be explored empirically in order to ascertain the most efficient and valid combination for assessing emotional and behavioral maladjustment. More research is needed to address these complex issues.

The Current Study

The research on multiple-gate screening procedures is quite minimal. Those studies that do exist oftentimes:

1. Focus on specific disorders rather than the assessment of general maladjustment (Hill et al., 2004; Lochman and The Conduct Problems Prevention Research Group, 1995; August et al., 1995; Power et al., 1998; August et al., 1992)
2. Utilize longer instruments or draw items from longer instruments rather than examining the utility of actual screening measures (Hill et al., 2004; Lochman and The Conduct Problems Prevention Research Group, 1995; August et al., 1995; August et al., 1992)

3. Fail to actually perform a multiple-gate procedure by simply adding measures into a regression equation (Hill et al., 2004; Lochman and The Conduct Problems Prevention Research Group, 1995)
4. Fail to tackle the issue of specific informant/gate combinations by systematically varying these factors ((Hill et al., 2004; Lochman and The Conduct Problems Prevention Research Group, 1995; August et al., 1995; August et al., 1992; Walker & Severson, 1994)

In the current study, we attempted to address a number of these limitations while focusing on four overarching research questions; First, are parents or teachers better as initial informants for screening for mental health problems of childhood? Second, are two gates better than a single gate? Third, are different informants better than just one informant in a two gate screening procedure? Finally, if a different informant is found to be superior to the same informant as the second gate, is a full BASC-2 necessary at the second gate or will a screener suffice?

The limited (multiple gate research) and contradictory (multiple informant research) findings cited previously, make it difficult to offer apriori hypotheses with confidence. With these limitations in mind we predicted that:

Both parent and teacher screeners would adequately discriminate between diagnosed and non-diagnosed children (Goodman et al., 2004). Second, we thought that the teacher screener would be superior to the parent screener as ADHD and EBD diagnoses are very relevant to the school environment and would therefore be brought to the teacher's attention. Additionally, Reynolds and Kamphaus' (1992) research with teacher rating scales has demonstrated that, on average, teacher ratings of child behavior

are more reliable than parent ratings at preschool, child, and adolescent age levels, and Loeber and colleagues (1991) found teacher reports of ADHD symptoms in elementary school children to be the best predictors of later impairment.

Third, and most tentatively, we predicted that adding a second gate would not be necessary. The teacher screener used in a prior investigation had a substantial coefficient alpha of .97, setting the stage for producing good validity evidence. Furthermore, Kamphaus and colleagues (2007) found almost identical predictive validity for behavioral and academic outcomes for the TRS-C screener and the full TRS-C BSI. Fourth, we predicted that use of multiple informants would substantially increase known- groups validity, as suggested by prior research indicating minimal correlations between teacher and parent reports (Achenbach et al., 1987). Other studies have documented the importance of multi-informant assessment (Verhulst, Dekker, & van der Ende, 1997; Power, Andrews, Eiraldi, Doherty, Ikeda, DuPaul, 1998; Jensen, Rubio-Stipec, Canino, Bird, Dulcan, Schwab-Stone, & Lahey, 1999; Kamphaus & Frick, 2002).

CHAPTER 2

METHODS

Sample

This study utilized a subsample of the Behavior Assessment System for Children - Second Edition (Reynolds & Kamphaus, 2004; BASC-2) general-population and clinical norm samples. The BASC-2 norming program took place at 375 U.S. sites in 257 communities and cities in 40 states (p. 113, Reynolds & Kamphaus, 2004) and represents a diverse sampling of the population by geographic region, SES, ethnicity, and child exceptionality based upon 2001 U.S. Census data. The sample has been screened for outliers including cases with high validity indexes. The items selected for all BASC-2 forms have been analyzed for child sex and racial/ethnic group item bias in the U.S. using differential item functioning procedures (Reynolds & Kamphaus, 2004).

The clinical norms sample was comprised of children and adolescents whose parents have identified them as having been diagnosed or classified with one or more emotional, behavioral, cognitive, or physical problems. Children in the clinical norm sample are not demographically matched to Census data since children with emotional and behavioral problems are not a random subset of the population.

The current study utilized two samples: a child and an adolescent sample. The child sample for the current study was comprised of 606 children (ages 6-11) who had both BASC-2 Parent Rating Scale – Child (PRS-C) version and Teacher Rating Scale – Child (TRS-C) version data available. This sample consisted of 302 (50%) males and

304 (50%) females. The socioeconomic status of the sample was derived based upon mother's education with 10% of the sample not completing high school, 40% of the sample having graduated from high school or received a GED, 31% completing three years in a college or technical school, and 21% completing four or more years of college or technical school. The sample was primarily Caucasian (61%). African Americans and Hispanics were represented in the sample as well, at 15% and 18%, respectively. There were relatively few Asian, American Indian, or "Other" students in the dataset (6% across the three categories).

This sample was further subdivided into two groups (diagnosed and control) to serve as outcomes for subsequent analyses. Parental report of diagnosis or special education classification was utilized as our outcome of interest. Parents completed a history form on which they indicated whether their child had any diagnoses or special education classifications. The diagnosed (DX) group consisted of 111 children (19% of sample) diagnosed either with Attention Deficit Hyperactivity Disorder (ADHD), an Emotional Behavioral Disorder (EBD), or both. Approximately 22% of the clinical sample was comorbid, having both ADHD and EBD diagnoses; we view the inclusion of children with comorbid diagnoses as a strong point of the sample since comorbidity tends to be the rule rather than the exception in child and adolescent mental health (Westen, Novotny, & Thompson-Brenner, 2004). Studies consistently find comorbidities in the range of 50% to 90% for Axis I disorders (Kessler, Stang, Wittchen, Stein, & Walters, 1999; Newman, Moffitt, Caspi, & Silva, 1998; Shea, Widiger, & Klein, 1992; Zimmerman, McDermut, & Mattia, 2000).

The DX sample constituted 19% of the total sample, which is consistent with findings that approximately 1 in 5 children have a diagnosable mental disorder (Campaign for Mental Health Reform, 2005). A control sample of children that did not carry any diagnosis according to parent reports served as the non-clinical sample (NC; N=495).

The adolescent sample was comprised of 716 adolescents (ages 12-18) who had both BASC-2 Parent Rating Scale – Adolescent (PRS-A) version and Teacher Rating Scale – Adolescent (TRS-A) version data available. This sample consisted of 378 (53%) males and 338 (47%) females. The socioeconomic status of the sample was derived based upon mother's education with 12% of the sample not completing high school, 37% of the sample having graduated from high school or received a GED, 30% completing three years in a college or technical school, and 18% completing four or more years of college or technical school. The sample was primarily Caucasian (70%). African Americans and Hispanics were represented in the sample as well, at 12% and 15%, respectively. There were relatively few Asian, American Indian, or "Other" students in the dataset (3% across the three categories).

This sample also was subdivided into two groups (DX and NC) to serve as outcomes for subsequent analyses. Once again, parental report of diagnosis or special education classification was utilized as our outcome of interest. The DX group consisted of 139 adolescents (20% of sample) diagnosed with either Attention Deficit Hyperactivity Disorder (ADHD), an Emotional Behavioral Disorder (EBD), or both. Approximately 29% of the clinical sample was comorbid, having both ADHD and EBD diagnoses.

Once again, the DX sample constituted 20% of the total sample consistent with findings population prevalence rates (Campaign for Mental Health Reform, 2005). A control sample of adolescents that did not carry any diagnosis according to parent reports served as the non-clinical sample (NC; N=577).

Instruments

TRS and PRS Screeners

The BASC-2 TRS and PRS child and adolescent screeners (Behavior Assessment System for Children – Second Edition Screening System, BASC-2 SS; Kamphaus & Reynolds, in press) were utilized as the first gate, and were derived from the 252 standardization items for the full BASC-2 TRS and PRS instruments using the general normative sample described previously. These screeners can be used with children and adolescents ages 6 to 18.

Test developers wanted item content to represent the major constructs of child adjustment (Reynolds & Kamphaus, 2004; Kamphaus & Reynolds, in press). Attending to these dimensions involved performing separate principal component analyses (PCA) for each factor-derived composite of the BASC-2 TRS and PRS (four and three dimensions respectively). This latent structure would include Internalizing Problems, Externalizing Problems, School Problems, and Adaptive Skills for the teacher screener, and Internalizing Problems, Externalizing Problems, and Adaptive Skills for the parent screener. Thus, the items with the highest factor loadings were selected within each dimension, with roughly equal representation of the dimensions. Additionally, in order to increase internal consistency to an alpha greater than .8 within the internalizing factor,

test developers added several internalizing items based upon content validity. Items were then combined to produce a total screener score (Kamphaus & Reynolds, in press).

PCA was utilized for item selection rather than principal factor analysis (PFA) in order to retain as much variance as possible in the screeners. As Fabrigar, Wegener, MacCallum, and Strahan (1999, p. 275) concluded "the objective of PCA is to determine the linear combinations of the measured variables that retain as much information from the original measured variables as possible." A criticism of the principal components extraction method is that it does not allow for measurement error in responses (i.e., the diagonal of the correlation matrix used with component extraction is set to 1.0, implying no measurement error in responses). However, the impact of the diagonal elements on the off diagonal elements is minimized as the number of items under study increases (Comrey & Lee, 1992). In developing the screeners, the sizeable number of items from which to draw will substantially lessen the impact of the diagonal elements.

The teacher screener consisted of 27 items: 6 externalizing (including Hyperactivity, Conduct Problems, and Aggression items), 9 internalizing (including Depression, Anxiety, and Somatization items), 6 school problems (including Attention Problems and Learning Problems), and 6 adaptive skills (including Functional Communication, Leadership, Social Skills, and Study Skills items) (see Table 3). Reliability estimates were high with a composite scale coefficient alpha of .944 and specific factor alphas ranging from .817 for Internalizing Problems to .932 for Externalizing Problems.

Similarly, the parent screener consisted of 30 items: 11 externalizing (including Hyperactivity, Conduct Problems, and Aggression items), 9 internalizing (including

Depression, Anxiety, and Somatization items), and 10 adaptive skills (including Attention Problems, Functional Communication, Leadership, Social Skills, and Activities of Daily Living items) (see Table 4). Reliability estimates were again high with a composite scale coefficient alpha of .930 and specific factor alphas ranging from .816 for Internalizing Problems to .882 for Adaptive Skills.

The BASC-2 TRS/PRS

The Behavior Assessment System for Children - Second Edition (Reynolds & Kamphaus, 2004; BASC-2), the successor to the original BASC (Reynolds & Kamphaus, 1992), consists of the Self-Report of Personality (SRP), Parent Rating Scales (PRS), and Teacher Rating Scales (TRS), a Student Observation System (SOS), and a structured developmental history form (SDH). The BASC-2 TRS and PRS were used as the second gates in our multiple gate screening procedure.

The BASC-2 TRS-C (ages 6-11) and TRS-A (ages 12-21) consist of 139 items, rated on a four-point response scale of frequency, ranging from “Never” to “Almost Always. These instruments have 15 scales and 5 composites including Adaptive Skills, Behavioral Symptoms Index, Externalizing Problems, Internalizing Problems, and School Problems. A four factor structure has been found through factor analysis. The Externalizing factor consists of Hyperactivity, Aggression, and Conduct Problems scales. The Internalizing factor consists of Anxiety, Depression, Somatization, Atypicality, and Withdrawal scales with the Adaptability scale having a secondary loading. The School Problems factor consists of Attention Problems and Learning Problems scales and the Adaptive factor is made up of the Adaptability, Social Skills, Leadership, Study Skills, and Functional Communication scales.

The BASC-2 PRS-C (ages 6-11) and PRS-A (ages 12-21) include 150 items rated on a four-point response scale of frequency, ranging from “Never” to “Almost Always. This instrument has 14 scales and four composites including Adaptive Skills, Behavioral Symptoms Index, Externalizing Problems, and Internalizing Problems. The BASC-2 PRS has been found to have a three factor structure through factor analysis. The Externalizing Problems factor consists of the Hyperactivity, Aggression, and Conduct Problems scales. The Internalizing Problems factor consists of Anxiety, Depression, Somatization, Atypicality and Withdrawal while the Adaptive Skills factor is made up of the Attention Problems, Adaptability, Social Skills, Leadership, Activities of Daily Living, and Functional Communication scales.

The BASC-2 manual (Reynolds & Kamphaus, 2004) provides three types of reliability evidence for the TRS and PRS: PRS median internal consistency (alpha) coefficients for the individual scales were reported at .84 for ages 6-7, .86 for ages 8-11, .85 for ages 12-14, and .85 for ages 15-18. Median TRS coefficient alphas for these same age groups were .88, .88, .88, and .86 respectively. Median test-retest reliability coefficients for the TRS were .88 for the child level and .79 for the adolescent level. The corresponding PRS test-retest coefficients were .84 and .82 (Reynolds & Kamphaus, 2004). The manual presents evidence of factor analytic support for the construct validity of the scales using both principal axis and covariance structure analysis methods. The TRS and PRS also exhibit high correlations with analogous scales from other behavior rating scales and systems (Reynolds & Kamphaus, 2004, 1992). Additionally, several independent reviews of the BASC have noted that the BASC TRS and PRS possess adequate to good evidence of reliability and validity support for a variety of inferences

using a variety of indicators (Doyle, Ostrander, Skare, Crosby, & August, 1997; Vaughn, Riccio, Hynd, & Hall, 1997).

Data Analyses

Variable properties and scaling were defined and coded as follows:

The BASC-2 subscales were expressed as general norming sample referenced T scores with M = 50 and SD = 10. Screener total scores were expressed as the total raw score for the sum of all items, where items were scored on a 0 to 3 scale. Classification was coded as dichotomous variable where 1 = diagnosis present and 0 = diagnosis not present.

Descriptive statistics were obtained for each variable including mean, median, and standard deviation. The typical checks for assumptions of normality and multicollinearity were not necessary as none of the analyses performed required these assumptions to be met. Also during this phase of the analysis, the dataset was checked for outliers using a cut score of greater than or equal to 3 standard deviations from the mean.

Again, this study focused on four overarching research questions: First, are parents or teachers better as initial informants for screening for mental health problems of childhood? Second, are two gates better than a single gate? Third, are different informants better than just one informant in a two gate screening procedure? Finally, if a different informant is found to be superior to the same informant as the second gate, is a full BASC-2 necessary at the second gate or will a screener suffice?

Specifically we:

1. Compared the known - groups validity of a teacher (TRS) and parent (PRS) screener (BASC-2 TRS screener versus PRS screener)

2. Assessed the need for a second gate in the screening procedure (1. TRS screener versus TRS screener + full BASC-2 TRS; 2. PRS screener versus PRS screener + full BASC-2 PRS; 3. TRS screener versus TRS screener + full BASC-2 PRS; 4. PRS screener versus PRS screener + full BASC-2 TRS)
3. Compared the known - groups validity of a two gate screening procedure using same versus different informant as the second gate (1. TRS screener + full BASC-2 TRS versus TRS screener + full BASC-2 PRS; 2. PRS screener + full BASC-2 PRS versus PRS screener + full BASC-2 TRS)
4. Compared the known - groups validity of a different informant, two gate screening procedure using a screener versus a full BASC-2 as the second gate (1. TRS screener + PRS screener versus TRS screener + full BASC-2 PRS; 2. PRS screener + TRS screener versus PRS screener + full BASC-2 TRS)

Question 1. BASC-2 SS PRS versus TRS screeners

We began by examining the known-groups validity of the newly developed, BASC-2 parent (PRS) and teacher (TRS) screeners. We also assessed whether the PRS or TRS screener was better at discriminating between the DX and NC groups. All analyses were performed for the PRS and TRS screeners on both the adolescent and child samples.

First, a Receiver Operating Characteristic (ROC) curve analysis was conducted in order to identify optimal cut-off scores for each screener. As described previously, the ROC curve assesses the effectiveness of a screening measure by evaluating the accuracy of discrimination between children with emotional, behavioral, or academic problems and those without. An area under the curve (AUC) of 1 defines a perfect test, while an area

of .5 represents a relatively worthless measure; ROC curve areas of .80-.90 are considered "good" discriminators while .90- 1 are considered "excellent." Other ROC curve analysis indices computed included overall hit rate, sensitivity, specificity, PPV, and NPV. Although a consensus does not exist as to the adequacy of various ROC curve analysis index values for a screening measure of emotional and behavioral adjustment, we have decided to use the most stringent standard identified in the literature of a minimum of .80 (Carran & Scott, 1992; Weis, Lovejoy, & Lundahl, 2005; Campbell, Bell, & Keith, 2001).

A logistic regression was performed using the selected screener cut-offs to code the predictor variable dichotomously (1 = positive screen; 0 = negative screen). Although collapsing a continuously scaled variable, such as screener score, into categories decreases variance, we chose to do so as this is how the screener would actually be utilized in practice. Ward statistics, Naglekerke's R^2, and odds ratios were used to determine the strength of the association between diagnosis and screener scores. The Wald statistic is a measure of significance of individual logistic regression coefficients, *B*, for the independent variable; higher values, in combination with degrees of freedom, indicate significance.

R^2 measures reflect the percentage of total variance in the dependent variable accounted for by the predictor variables; however, with a dichotomous variable such as "diagnosis" we have inherent heteroscedasticity with a different error variance for each value of the predicted score. Therefore, each value would have a different measure of variance accounted for, preventing us from examining variance accounted for in the universal sense for logistic regression. Naglekerke's R^2 can be viewed as a measure of the

strength of association and an approximation to Ordinary Least Squares (OLS) R^2. This value is a version of the Cox and Snell index that has been adjusted so that the maximum value it can attain is a 1.00. Additionally, the logistic R^2 measures for good logistic regression models are generally smaller than R^2 for good models in OLS (Cohen, Cohen, West, & Aiken, 2003). Although we considered Naglekerke's R^2 when assessing the validity of the screeners, this index does not warrant substantial attention due to its inherent difficulty of interpretation. As Pedhazur (1997, p.758) explained, "…even authors who advocate the use of R^2 as an index of fit concede that it is of little utility for ascertaining the effects of independent variables on the dependent variable."

Therefore, our comparison of logistic regression results for the screeners focused on odds ratios. An odds ratio tells us by what amount the odds of being in the case group (DX) are multiplied when the predictor is incremented by one unit. In other words, the odds ratio tells us the odds of being diagnosed given a positive screen (Pehazur, 1997). An odds ratio greater than 1 indicates that the odds of a positive- screen case having a diagnosis is greater than the odds of a negative- screen case having a diagnosis.

We then compared the results of these analyses for the PRS screener and the TRS screener in order to assess whether one screener was superior. Logistic regression indices were examined with a focus on odds ratios. In comparing ROC curves, Hanley and McNeil's (1983) method for comparing AUCs was used. Additionally, overall hit rates, sensitivity, specificity, PPV, and NPV values were compared using a two-tailed dependent samples Z - test of proportions at the .05 probability level.

Question 2. Single gate versus Two gates

Using the full BASC-2 PRS-C and TRS-C for the child sample and the full BASC-2 PRS-A and TRS-A for the adolescent sample, we then evaluated whether adding a second, more comprehensive gate, either same or different informant, significantly increased the classification accuracy of the first gate screener. Therefore, for the same informant condition, we used the full PRS as the second gate for the sample identified using the PRS screener, and for the sample identified by the TRS screener, we used the full TRS. Then for the different informant condition, we used the full TRS as the second gate for the sample identified using the PRS screener, and for the sample identified by the TRS screener, we used the full PRS.

In order to conduct an authentic multiple gate procedure, we applied the second gate, full BASC-2, criterion on the sample of children identified by the screener in the previous step. BASC-2 cut-off scores of 60, 65, and 70 or above on any of the BASC-2 clinical subscales and scores of 30, 35, 40 or less on any adaptive subscale (see Methods for descriptions of subscales) were examined using ROC curve indices to identify an optimal cut-off score for the second gate, full BASC-2, measure. We then compared the number of false positives before the second gate was applied to the number of false positives after the second gate was applied to determine whether adding a second gate increased classification accuracy by significantly decreasing the number of false positives. We also considered the number of false negatives created in applying a second gate.

We then examined odds ratios at the second gate. If most of the prediction was done at the first gate then we would expect odds ratios to be quite small; however, if the second gate is needed, then these indices should be greater than one.

Question 3. Second Gate: Same versus Different Informant

In this section, we assessed whether different informants discriminated better than the same informant when utilizing a two gate screening procedure. We compared the overall hit rate, sensitivity, specificity, PPV, NPV, number of false positives, and number of false negatives obtained when using a different informant than the first gate informant as the second gate to those obtained when using the same informant as the second gate. Once again, we considered odds ratios at the second gate using logistic regression.

Question 4. Second Gate, Different Informant: Screener versus Full BASC-2

Finally, we examined whether a brief screener might be utilized as the second gate of a different informant, two gate screening procedure. We did so by assessing the known - groups validity of a different informant, two gate screening procedure using a screener versus the full BASC-2 as the second gate. We began by examining the decrease in false positives when adding a second gate screener to the first gate screener in order to assess the overall utility of having a second gate screener versus just a first gate. Then, when using the TRS screener as the first gate, we compared the odds ratios, sensitivity, specificity, PPV, NPV, and overall hit rate of the full BASC-2 PRS as the second gate to the PRS screener as the second gate. When using the PRS screener as the first gate, we compared the classification accuracy of using the full BASC-2 TRS to using the TRS screener at the second gate.

CHAPTER 3

RESULTS

Preliminary Statistical Analyses

Data screening identified several significant multivariate outliers using the aforementioned 3 standard deviation obtained score criterion. The scores for these cases were examined manually and found to be within valid ranges. As there was no theoretical rationale for excluding them, all cases with complete data were included in the final analyses.

Question 1. BASC-2 SS PRS versus TRS screeners

ROC Curve Analysis

Four ROC curve analyses were performed: TRS screener child sample, TRS screener adolescent sample, PRS screener child sample, and PRS screener adolescent sample. Cut-off selection ended up being quite apparent, as ROC curve analyses usually produced one or two superior cut-off scores with adjacent score alternatives exhibiting an obvious decrease in sensitivity or specificity values. In deciding optimal cut scores for each screener, we attempted to use the .80 criterion for balancing sensitivity and specificity. When choosing between two cut-off scores, we sacrificed specificity for sensitivity in order to minimize false negatives. Thus our first requirement was to keep sensitivity above .80, followed by identifying the optimal specificity given this requirement.

False positives are more acceptable than false negatives for a first gate screening instrument as we want to catch as many children with emotional and behavioral problems as possible at this stage. Children with emotional and behavioral problems who are missed at this gate are not recoverable through later assessment or "lost for good." False positive errors, on the other hand, can be corrected through the addition of later gates.

The Area Under the Curve (AUC) for the TRS screener on the child sample was .863 indicating a good screening test under the parameters described previously (see Figure 1). The optimal cut off score was identified as a raw score of 27, yielding a sensitivity value of .811 and specificity of .764 (see Table 5). This cut-off was selected because it produced the best specificity value when keeping sensitivity above .80. PPV and NPV were computed and found to be .435 and .947 respectively. The overall hit rate was .772 (see Table 6).

For the adolescent sample, the Area Under the Curve (AUC) for the TRS screener was .829, once again indicating a good test under the desired parameters (see Figure 2). The optimal cut off score was found to be a raw score of 20, yielding a sensitivity of .820 and specificity of .688 (see Table 7). This cut off allowed for the highest specificity without forcing sensitivity to drop below .80. PPV and NPV were computed and found to be .388 and .941 respectively. The overall hit rate was .714 (see Table 8).

The Area Under the Curve (AUC) for the PRS screener on the child sample was .891 indicating a good test under the parameters described previously (see Figure 3). The optimal cut off score was found to be 33, yielding a sensitivity of .838 and specificity of .812 (see Table 9). This was one of two possible cut-offs, 33 and 34, with both sensitivity and specificity above .80. We selected the cut-off with the higher sensitivity,

consistent with our decision to favor sensitivity over specificity when necessary. PPV and NPV were computed and found to be .500 and .957 respectively. The overall hit rate was .817 (see Table 10).

Lastly, the Area Under the Curve (AUC) for the PRS screener on the adolescent sample was comparable to that of the other screeners at .881 (see Figure 4). The optimal cut off score was found to be 30, yielding a sensitivity of .813 and specificity of .775 (see Table 11). Once again, this cut-off was selected because it produced the best specificity value when keeping sensitivity above .80. PPV and NPV were computed and found to be .465 and .945 respectively. The overall hit rate was .782 (see Table 12).

Logistic Regression

Logistic regression results (see Table 13) produced Naglekerke's R^2 values ranging from .251 for the teacher screener on the adolescent sample to .400 for the parent screener on the child sample. All Ward statistics were found to be significant at the .0001 level. Odds ratios ranged from 10.06 for the teacher screener adolescent sample to 22.33 for parent screener child sample. In other words, these results indicated that the children in our sample were anywhere from 10.06 to 22.33 times more likely to be diagnosed given a positive screen.

Comparison of Teacher and Parent Screeners

In comparing ROC curves, Hanley and McNeil's (1983) method for comparing AUCs was used. For the adolescent sample, the parent screener AUC was found to be superior at a statistically significant level ($Z=2.63$, $p<.05$). In contrast, the difference between the parent and teacher screener AUCs for the child sample were not statistically significant (see Table 13).

ROC curve indices, including overall hit rate, sensitivity, specificity, PPV, and NPV values, were compared for the parent and teacher screeners using a two-tailed Z test of proportions. For both the child and adolescent samples, the parent screeners had superior specificities ($Z=2.05$, $p<.05$, $Z=3.73$, $p<.05$ respectively) and PPVs ($Z=2.27$, $p<.05$, $Z=2.96$, $p<.05$ respectively) at statistically significant levels. For the adolescent sample, the parent screener also had a higher overall hit rate ($Z=2.97$, $p<.05$) that was statistically significant. Odds ratios were higher for the parent screeners at 22.33 versus 13.85 for the child sample and 14.94 versus 10.05 for the adolescent sample (see Table 13).

Question 2. Single Gate versus Two Gates

This set of analyses assessed the utility of adding a full BASC-2 second gate to our first gate screeners. We examined this question using the same informant as the second gate as well as a different informant; however, the focus was on comparing classification accuracy of a single gate versus a two gate procedure, regardless of informant.

False positives, false negatives, sensitivity, specificity, PPV, NPV, and overall hit rate were compared across the three proposed BASC-2 cut scores and a BASC-2 cut score of 70 and above on clinical scales and 30 and below on adaptive scales was found to be optimal (see Table 14). The 70/30 cut score was found to have superior overall hit rates for all screeners and samples. Additionally, the two lower cut score alternatives tended to result in unacceptably low specificity, indicating very high false positive rates.

For the second gate, false positives and false negatives should be given equal consideration as opposed to when considering cut-off criteria for the first gate screener

where the goal is to minimize false negatives. Although the number of false negatives increased as cut-off scores became more extreme, the decrease in false positives was more substantial than the increase in false negatives, causing us to select the aforementioned cut-off criteria.

The 70/30 cut score selection also made clinical sense as two standard deviations above/below the mean has been a common clinical standard for decades and was suggested by the BASC-2 manual as indicative of clinical significance (Reynolds & Kamphaus, 2004). Since our outcome was "diagnosis," clinical significance would be the most accurate way to correctly classify these cases.

We first examined the utility of adding a same-informant second gate. For the child sample, adding a BASC-2 TRS-C second gate to the TRS screener decreased false positives by 39%, and adding a BASC-2 PRS-C second gate to the PRS screener decreased false positives by 47%. For the adolescent sample, adding a BASC-2 TRS-A second gate to the TRS screener decreased false positives by 57% and adding a BASC-2 PRS-A second gate to the PRS screener decreased false positives by 49% (see Table 15).

We then examined the utility of adding a different-informant second gate. For the child sample, adding a BASC-2 PRS-C different-informant second gate to the TRS screener decreased false positives by 75% and adding a BASC-2 TRS-C different-informant second gate to the PRS screener decreased false positives by 72%. For the adolescent sample, adding a BASC-2 PRS-A second gate to the TRS screener decreased false positives by 74% and adding a BASC-2 TRS-A second gate to the PRS screener decreased false positives by 72% (see Table 15).

For the same informant second gate, false negatives ranged from 12 cases for the parent screener child sample to 32 cases for the teacher screener adolescent sample. Total number of false negatives across both gates ranged from 30 for the parent screener child sample to 57 for the teacher screener adolescent sample. When using a different informant, false negatives ranged from 20 cases for the teacher screener child sample to 36 cases for the parent screener adolescent sample with total number of false negatives ranging from 41 (teacher screener child sample) to 62 (parent screener adolescent sample) (see Table 15).

When using the same informant as a second gate, odds ratios ranged from 2.13 for the teacher screener, child sample to 6.061 for the parent screener, child sample, indicating that the cases were anywhere from 2.13 to 6.061 times more likely to be diagnosed given a positive second gate screen. Odds ratios ranged from 4.912 for parent screener child sample to 10.62 for teacher screener child sample when a different informant was used as the second gate (see Table 16). Therefore, results indicated a clinically meaningful improvement in classification accuracy when adding a second gate, either same or different informant, to the screener.

Question 3. Second Gate: Same versus Different Informant

The previous section indicated the importance of having a second gate and suggested the utility of using a different informant in this second gate. In this section, we focused on directly comparing the classification accuracy of a same informant second gate to a different informant second gate. By using a different informant as the second gate, we decreased false positives by 59% more than using the same informant for the teacher screener and 47% more for the parent screener. By using a different informant as

the second gate in the adolescent sample, we decreased false positives by 39% more than using the same informant for the teacher screener and 46% more for the parent screener (see Table 17).

When the teacher screener was the first gate, overall hit rate, specificity, PPV, and NPV were all found to be statistically superior when using a parent as the second gate as opposed to the same informant (teacher) for both the child (Overall Hit Rate: $Z=4.55$, $p<.05$; Specificity: $Z=7.92$, $p<.05$; PPV: $Z=4.55$, $p<.05$; NPV: $Z=3.05$, $p<.05$) and adolescent (Overall Hit Rate: $Z=3.34$, $p<.05$; Specificity: $Z=4.33$, $p<.05$; PPV: $Z=3.53$, $p<.05$; NPV: $Z=2.42$, $p<.05$) samples (see Table 18).

For the parent screener child and adolescent samples, there was an increase in false negatives when using a different informant when compared to the number of false negatives when using the same informant as a second gate; however, this increase in false negatives was only found to be statistically significant, as evidenced by the NPV, for the child sample ($Z=2.39$, $p<.05$). Specificity ($Z=5.02$, $p<.05$) for this sample was found to be statistically significant in favor of the different-informant procedure (see Table 18).

For the parent screener adolescent sample, sensitivity was found to be statistically higher in the same informant procedure ($Z=2.94$, $p<.05$). On the other hand, specificity ($Z=5.37$, $p<.05$) and PPVs ($Z=2.39$, $p<.05$) were found to be superior in the different informant procedure at a statistically significant level. NPVs were not statistically different for the same and different informant procedures in this sample (see Table 18). All odds ratios were higher for the multiple-informant procedure with the exception of the parent screener child sample which had an odds ratio of 6.061 for the same informant

procedure and 4.912 for multiple informant procedure. Therefore, results generally supported the utility of using a different informant as the second gate (see Table 16).

Question 4. Second Gate, Different Informant: Screener versus Full BASC-2

As our results have indicated thus far that two gates are better than one gate and different informants better than the same informant, we then considered whether the use of a different informant, second gate "screener" would be adequate as opposed to making informants complete an entire BASC-2 measure as the second gate. Adding a second gate screener was found to be superior to relying only on a first gate as evidenced by the large decrease in false positives for all samples and screeners, ranging from 49% to 68% (see Table 19).

We then compared the classification accuracy of using a screener versus a full BASC-2 as the second gate. For both the child and adolescent samples, results indicated that sensitivity was superior to a statistically significant extent when using the screener as the second gate regardless of screener order (Gate 1: parent, Gate 2: teacher or vice versa). When the parent screener was used as a first gate, NPV was also found to be statistically higher when the TRS screener was used as the second gate rather than the full BASC-2 TRS in both child ($Z=2.73$, p,.005) and adolescent ($Z=2.83$, p,.005) samples; however, specificity was found to be statistically higher when using a full BASC-2 TRS as the second gate for both child ($Z=2.42$, $p<.005$) and adolescent ($Z=5.53$, $p<.005$) samples. For the adolescent sample, using the full BASC-2 PRS as the second gate was also found to be statistically superior ($Z=2.91$, $p<.005$) when the TRS screener was used as the first gate (see Table 20).

CHAPTER 4

DISCUSSION

Fortunately, the results of our investigation of the known - groups validity of parent and teacher screeners of children's mental health are unusually clear, as are implications for screening programs. These results suggest that the new BASC-2 PRS and TRS screeners are promising as gate one screening instruments, parents appear to do a better job than teachers as first gate informants, and adding a comprehensive behavior rating scale as second gate significantly improves identification accuracy. Additionally, utilizing a different informant at the second gate generally improves identification accuracy further. Lastly, when implementing a different informant, two gate screening procedure, a screener is a valid option as the second gate in place of a longer behavior rating scale such as the full BASC-2 depending on the purpose of screening. The details of these findings and their juxtaposition in the larger research literature are described in the next sections.

Question 1. BASC-2 SS PRS versus TRS screeners

Although several indices were below our stringent .80 criterion, overall the screener index values are very promising for a first gate screening measure, especially when compared to the results of other research studies utilizing these indices to evaluate classification accuracy of instruments (Petras, Howard, Chilcoat, Leaf, Ialongo, & Kellam, 2004; Hill et al., 2004; Power et al., 1998; Campbell et al., 2001; Jellinek, Murphy, & Burns, 1986; Goodman et al., 2003; Gardner et al., 1999; Leon et al., 1999;

Schmitz, Kruse, Heckrath, Alberti, & Tress, 1999; Bennett, Lipman, Brown, Racine, Boyle, & Offord, 1999; Bennett & Offord, 2001); however, it must be noted that these studies all utilized different criterion variables thus making conclusions based upon direct comparison of indices tentative at best.

For example, Schmitz and colleagues (1999) concluded that a mental health screening instrument, the General Health Questionnaire-12, with a sensitivity of .68, specificity of .65, PPV of .53 and NPV of .78 "showed acceptable qualities for diagnosing mental health disorders in the primary care sector. The use of the GHQ-12, employed as first step, supplemented by a second stage interview, may enhance the detection rate of mental disorder…(p. 365)." Lueng and colleagues (2005) examined the psychometric properties of the new parent DPS- 4.32 version using a community sample (N=541) of Chinese children, and concluded that the measure holds great promise as a screener due to its sensitivity (.68), specificity (.91), and NPV (.98) even though PPV was found to be .34. Moreover, Leon and colleagues (1999) found that a depression screen with PPVs of .48 and .38, for depression and panic disorder respectively, compared favorably with primary care screens used for mental (Burnam, Wells, Leake, & Landsverk, 1988) and medical disorders (Soost, Lange, Lehmacher, & Ruffing-Kullman, 1991; Sickles, Ominsky, Sollitto, Gavlin, & Monticciolo, 1990; Littrup, Lee, & Mettlin, 1992).

The screeners also met our objective of minimizing false negatives, as indicated by the high NPVs (ranging from .941 to .957), which is crucial in a first gate screening measure. The larger number of false positives (as indicated by lower PPVs ranging from .388 to .500), is more acceptable for a first gate screening instrument as false positive

errors can be corrected through the addition of later gates; children with emotional and behavioral problems who are missed at this gate, on the other hand, are not recoverable through later assessment.

Additionally, our odds ratios, ranging from 10.06 to 22.33, were found to be remarkable large within the context of the literature. Petras and colleagues (2004) found odds ratios ranging from 1.37 to 2.05 when using the TOCA-R to predict later violence in adolescent males. In 2005, Petras and colleagues found odds ratios ranging from 1.56 to 2.48 in a sample of adolescent females. Timbremont, Braet, and Dreesen (2004) found a 1.36-point increase in risk for depressive disorder associated with each 1 point increase in a Children's Depression Inventory (CDI) score. Rettew, Copeland, Stanger, and Hudziak (2004) investigated the accuracy of the Junior Temperament and Character Inventory (JTCI) in predicting a number of disorders and found odds ratios ranging from .49 for ADHD to 1.38 for disruptive behavior disorders. Once again, differences in criterion measures must be kept in mind. For example, Petras and colleagues (2004) were seeking to predict *later* violence while our study focused on already diagnosed cases.

The results of this particular study suggest that the parent screeners were superior in both the child and adolescent samples. For both samples, these screeners had higher odds ratios as well as superior specificities and PPVs at a statistically significant level. For the adolescent sample, the parent screener also had statistically significantly higher overall hit rate. Therefore, the parent screeners were able to correctly identify a higher proportion of individuals without diagnoses (specificity=true negatives). Additionally, a higher proportion of cases with positive screens on the parent screener actually had

diagnoses (PPV). Both of these indices suggest that the parent screeners produced a lower number of false positives than the teacher screeners. However, we propose that both parent and teacher screeners appear to be generally effective.

Question 2. Single Gate versus Two Gates

First, we find it necessary to point out the inherent difficulty of assessing the utility of adding a second gate. By applying the second gate measure only to those cases identified by the first gate, we have significantly changed the sample. Many of the true negatives as well as some false negatives are eliminated after the first gate screening. Therefore, the sample to which the second gate is applied has many more diagnosed cases than the original sample.

ROC curve indices are significantly affected by the characteristics of the remaining sample. For example, specificity is calculated by dividing the number of true negatives by all non-diagnosed cases (true negatives + false positives). The majority of true negatives were eliminated at Gate 1 while the false positives from the Gate 1 screening remain in the Gate 2 sample thus causing specificity to decrease. Therefore, we cannot directly compare ROC curve indices between the single gate and two gate screening procedures. Additionally, false negatives cannot be directly compared because the false negatives from the first gate screening are not part of the second gate screening sample. However, we can look at the raw number of false positives in order to assess whether a second gate improves classification accuracy to a clinically meaningful extent.

Our results showed significant improvements in classification accuracy when adding a second gate, either same or different informant, to the screener as evidenced by large decreases in false positives, ranging from 39% to 57% when using the same

informant, and 72% to 75% when using a different informant. Odds ratios ranged from 2.13 to 10.62 indicating that the cases were anywhere from 2.13 to 10.62 times more likely to be diagnosed given a positive second gate screen. The TRS screener child sample using the same informant as Gate 2 had the least impressive odds ratio; however, they were still 2.13 times more likely to be diagnosed, a meaningful difference in the context of the literature (Petras et al., 2004; Petras, Ialongo, Lambert, Barrueco, Schaeffer, Chilcoat, & Kellam, 2005; Timbremont, Braet, & Dreesen, 2004), and the second gate decreased false positives by 39%, thus indicating a meaningful increase in classification accuracy.

Total false negatives after the second gate ranged from 30 (PRS screener, child sample) to 57 (TRS screener, adolescent sample) for same informant and 41 (TRS screener, child sample) to 62 (PRS screener, adolescent sample) for different informant. False positives ranged from 49 (PRS screener, child sample) to 77 (TRS screener, adolescent sample) for same informant and 26 (PRS screener, child sample) to 47 (TRS screener, adolescent sample) for different informant. These findings indicated a relatively small number of misclassified cases following the second gate screening.

Moreover, as stated previously, at the second gate false positives and false negatives should be given equal consideration as opposed to the first gate screening where the goal is to minimize false negatives. Positively screened cases at this gate may be referred for more comprehensive assessment or intensive interventions, requiring more resources at this gate thus making false positives less desirable. Additionally, the false negatives at this second gate would hopefully still be receiving some intervention due to

being identified at Gate 1. These cases are already in the system and not "lost for good" like false negatives at the first gate.

Question 3. Second Gate: Same versus Different Informant

Results from our study also indicated that using a different informant as the second gate generally resulted in improved classification accuracy as compared to using the same informant as the second gate. For all screeners and samples, the number of false positives was drastically decreased when using a different informant as opposed to the same informant as the second gate.

When the TRS screener was the first gate, the benefit of using a different informant as the second gate was clear as overall hit rate, specificity, PPV, and NPV were all found to be statistically superior when using a parent as the second gate as opposed to the same informant (teacher) for both the child and adolescent samples.

Although there was an increase in false negatives for the PRS screener in the different informant condition when compared to the number of false negatives when using the same informant as a second gate, this increase in false negatives was only found to be statistically significant for the child sample, and specificity for this sample was found to be statistically significant in favor of the different-informant procedure indicating a large improvement in correctly identifying individuals without diagnoses (true negatives).

For the PRS screener adolescent sample, sensitivity was found to be statistically superior in the same informant procedure, suggesting that the same informant procedure was better at correctly identifying diagnosed cases; however, specificity and PPVs were found to be superior to a statistically significant extent in the different informant

procedure, indicating that the using the BASC-2 TRS-A as the second gate resulted in improved identification of non-diagnosed cases. Additionally, a higher proportion of cases with positive screens on the different informant procedure actually had diagnoses (PPV). Both of these indices suggested that the different informant (PRS screener, BASC-2 TRS-A second gate) procedure produced superior classification results, that is a lower number of false positives than the same informant procedure (parent at both gates).

All odds ratios were higher for the multiple-informant procedure with the exception of the PRS screener, child sample which had an odds ratio of 6.061 for the same informant procedure and 4.912 for multiple informant procedure. Despite lack of improvement in odds ratios, the PRS screener, child sample did appear to benefit from having the teacher as the second gate as evidenced by the 47% greater decrease in false positives.

Question 4. Second Gate, Different Informant: Screener versus Full BASC-2

In considering whether the use of a different informant, second gate "screener" would be adequate as opposed to making informants complete an entire BASC-2 measure as the second gate, we began by assessing whether adding a second gate screener was superior to relying only on a first gate. Results supported the utility of adding a second gate screener, as evidenced by the large decrease in false positives for all samples and screeners, ranging from 49% to 68%.

We then directly compared the classification accuracy of using a screener versus a full BASC-2 as the second gate. In general, results indicated that the screener second gate resulted in superior sensitivity thus minimizing false negatives while using the full

BASC-2 resulted in superior specificity and fewer false positives. Overall hit rates were found to be comparable.

These results suggest that employing a screener as the second gate would be a valid option depending on the purpose of screening and the particular service delivery model being implemented. As results indicated, using the screener as the second gate appears to minimize false negatives while using the full BASC-2 minimizes false positives. Therefore, if the goal is to identify every child who is at risk for emotional and behavioral problems, then the screener would be preferred; however, if the goal is to ensure that resources are not wasted on children without problems and to minimize un-necessary stress and stigma for the misidentified child and his/her family then the full BASC-2 is superior.

Additionally, by using a screener as opposed to a full behavior rating scale as the second gate, we are decreasing the time and effort needed, but also decreasing the amount of information obtained. Therefore, if the purpose of screening is to identify children who will receive full evaluations, then using two screeners would be adequate. However, if the purpose of screening is to identify children for the purpose of providing interventions, then the extra information provided by a full behavior rating scale might be desired. Therefore second gate selection should be guided by practical concerns such as purpose of screening, resources available, and service delivery model.

Limitations

Outcomes

First, we must consider the limitation of our criterion (EBD/ADHD diagnosis). These children are already diagnosed with disorders, suggesting some level of severity.

However, the purpose of the screener is to identify children early, BEFORE risk becomes a more serious, diagnosable problem. A number of false positives still remained after the second gate, ranging from 49 (parent child) to 77 (teacher adolescent) when using same informant and 26 (parent child) to 47 (teacher adolescent) when using a different informant as the second gate; however, some of the children identified as false positives in this study may actually be children with subsyndromal behavioral and emotional problems. For example, Leon and colleagues (1999) found that a substantial portion of those cases labeled as false positives, had significantly more psychopathology and functional impairment than those with true negative results. Longitudinal studies would be necessary in order to ascertain whether the screeners can identify children "at-risk" and are actually "predictive" of future emotional and behavioral problems. The fact that our study was concurrent rather than predictive must be kept in mind when comparing it to other studies as well.

Additionally, the fact that our criterion of interest (diagnosis/classification) was based upon parent report is a potential limitation. First, we do not have adequate documentation of the reliability and validity of the parent's report. Moreover, parent knowledge and reporting of diagnosis/classification may have affected their ratings on the BASC-2 screener, thus possibly contributing to the finding that the parent screener performed better than the teacher screener as a first gate measure.

Cut scores

Although the cut scores we selected for the first gate and second gate screenings were optimal for our study, these cut scores cannot necessarily be generalized for several reasons. First, these cut-offs are sample specific. Sensitivities and specificities of cut

scores would change depending on the sample to which they are applied. Additionally, as explained above, our outcome of interest was "diagnosis." Therefore, our cut scores were selected to optimize ability of the screener to discriminate between those with diagnoses and those without. For example, we selected our second-gate cut score of 70/30 in order to identify cases at clinically significant levels. However, if one were to want to identify children at risk for emotional and behavioral problems than a 60/40 cut score may have been more appropriate.

Cut score selection depends on the goal and/or purpose for screening. By altering cut scores, sensitivity and specificity may be increased or decreased accordingly. For example, the number of false negatives may be decreased by lowering the cut - score; however, this would cause an increase in false positives. The desired balance between these two errors may differ depending on a number of factors including purpose of screening as well as financial and personnel resources. False negative errors result in children with emotional and behavioral problems being missed and denied necessary services; therefore, if the purpose of you measure is to catch ALL children with emotional and behavioral problems, false negatives should be minimized through the selection of lower cut scores. However, false positives create difficulties in terms of finances, time, and personnel. By serving children who do not necessarily need services, valuable resources are wasted. False positives can also result in un-necessary stress and stigma for the misidentified child and his/her family.

Another issue that will have to be discussed and addressed is the fact that we found different cut scores for our child and adolescent samples. Therefore, although both children (ages 6-11) and adolescents (ages 12-18) will be given the same screeners, they

may need to have different cut-scores depending on age in order to ensure optimal classification accuracy.

Sample

The sample for the current study required both parent and teacher ratings, thus resulting in a self-selected subsample of the full BASC-2 norming sample. However, the sample does appear to be representative of the overall normative sample as demonstrated through demographic information presented in the Methods section.

BASC-2 specific

Lastly, we wish to remind readers that this study is focused on the BASC-2, resulting in an item pool that has been pre-defined. The BASC-2 does cover the majority of domains of emotional and behavior adjustment that have been found to be important in childhood and adolescence (Reynolds & Kamphaus, 2004); however results cannot necessarily be generalized to all multiple gate procedures. More research should be done using other instruments and rating scales in order to determine the generalizability of results.

Directions for Future Research

Generalizing Findings

As noted above, more screening research must be done using other instruments, behavior rating scales, and outcomes in order to determine the generalizability of results of the current study. Longitudinal studies of multiple gate screening procedures are particularly crucial for determining whether screeners can actually identify children who are at-risk for and later develop significant behavioral and emotional problems.

In reviewing the literature, researchers have utilized a number of outcome measures in the screener validation process. August and colleagues (1992) focused on functional impairment using measures of behavioral (using the Internalizing and Externalizing dimensions of the CBCL-PRF; Achenbach, 1991), social (using various subscales of the Walker-McConnell Scale of Social Competence and School Adjustment; Walker & McConnell, 1995), and academic (using Woodcock-Johnson Tests of Achievement; McGrew & Woodcock, 2001) adjustment. In this way, they could assess impairment independent of clinical diagnoses. An individual found to have functional impairments in any of these three areas would be positively identified when performing the ROC curve analysis.

Other possible criterion/outcome measures include other DSM-IV (APA, 1994) or ICD (WHO, 1993) diagnoses (Tarnopolsky, Hand, McLean, Roberts & Wiggins, 1979; Winter, Steer, Jones-Hicks, & Beck, 1999), special education classifications, longer, validated parent and teacher behavior rating scales (Simonian & Tarnowski, 2001), mental health referral and treatment histories (Saunders & Wojcik, 2004), clinician or teacher-rated levels of impairment (Kelleher, Moore, Childs, Angelilli, & Comer, 1999; Pagano et al., 2000; Saunders & Wojcik, 2004), as well as diagnostic structured interviews such as the Structured Clinical Interview for DSM-IV (SCID-IV; Kobak, Taylor, Dottl, Greist, Jefferson, Burroughs, Katzelnick, & Mandell, 1997; Leon et al., 1999; Pagano et al., 2000; Schmitz et al., 1999).

As McFall (2005, p. 318) explained, “Only when both sides of the assessment equation have been nailed down is it possible to evaluate what, if anything, the total assessment effort has revealed. Unfortunately, criterion assessment has not received the

attention to date that it requires." No "gold standard" presently exists in psychological assessment research (August et al., 1992; McFall, 2005) and all commonly used criterion measures described previously have significant limitations.

For example, obtaining teacher's ratings of students often leads one to suspect method variance since teachers are usually the respondent on the screening measure as well. Classroom observations of behavior are often utilized as outcome variables; however, observations only allow a limited sampling of behavior and have therefore been found to be less valid than teacher ratings for classification and diagnostic purposes (Lett & Kamphaus, 1997). Special education placement, another commonly used outcome measure, is of unknown reliability and validity and has been found to be determined by factors other than a child's academic performance or behavior in school, including a child's sex or race/ethnicity. Kim and Rowe (2004) found that children in special education and those in regular education had identical teacher ratings of their behavior, thus raising the question of why one child was "placed" and another was not (Kim & Rowe, 2004). Lastly, the use of DSM diagnosis as an outcome measure is common practice in psychological assessment literature; however, many of the diagnostic categories found in the DSM have yielded inter-diagnostician reliability estimates lower than the internal consistency estimate of most psychological assessment instruments (i.e. .97; Kamphaus & Frick, 2002).

Due to the limitations of all outcomes measures, researchers must emphasize the need for replication when conducting this type of research. Utilization of a "bootstrapping" approach would suggest that we must continually validate new measures against known inferior measures until enough evidence is accumulated to demonstrate

that the new measure is superior. In this regard, Schmidt, Kotov, and Joiner (2004) have observed, "…philosophers of science have shown that it is possible to start with fallible indicators and gradually improve on them, simultaneously refining assessment of the construct (Meehl, 1992, p. 141)."

Assessing the Cost/Benefit Ratio

In addition to accuracy of identification, the practicality of the instrument is of the utmost importance. One must balance the amount of information needed to reliably identify those children at-risk for emotional and behavioral problems against the time and monetary limitations of those who will be utilizing the instrument. The scientific literature indicates that the impracticality of many screening measures has largely contributed to their lack of adoption on a wide scale in both pediatric and school settings (Flanagan, Bierman, & Kam, 2003; Saunders & Wojcik, 2004; Schmitz, Kruse, Heckrath, Alberti, & Tress, 1999). Therefore, one must compare the time and cost of adding additional levels of assessment and informants against the increase in time and cost that will result from the elimination of a level or informant. For example, when a level is eliminated, the number of children receiving full diagnostic assessments will increase due to the decrease in children being removed from consideration by that level of assessment.

We must assess what schools and primary care settings will tolerate in terms of time and costs. Both teachers and physicians are extremely busy and their time is valuable. If the screening process takes too long for teachers to complete, universal screening is unlikely to be adopted successfully. Moreover, the costs associated with implementing a universal emotional and behavioral screening program, including personnel and materials costs, must be explored. One must balance the amount of time

and money it will take to complete the screener against the amount of information needed to make accurate predictions.

In general, we must determine whether a universal screening program adds to the burdens that teachers and physicians already face or alleviates burdens of financial expense and time by increasing accuracy of referral and identifying children earlier. Presumably, accurate screening will decrease the need for time- and money- consuming procedures such as special education referral and full evaluations as well as more intensive interventions.

Assessing Consequential Validity

Ultimately, evidence of validity based on the consequences of screening will be necessary to defend its use. Early identification must lead to better behavioral and emotional outcomes for children; better than would be expected in the case of current typical service identification practices. Research must be done to assess whether the intended consequences of such a program come to fruition. In order to do this, it would be necessary to implement a schoolwide or even districtwide screening program and evaluate the actual consequences longitudinally across a number of years.

For example, we would want to determine whether the screener was being utilized as intended. Due to the ease of administration, schools may be tempted to use the screener as a diagnostic tool rather than as an indication of possible risk. Screeners are inherently less broad-based, assessing a necessarily limited range of behaviors and emotions due to their shorter length. Moreover, in choosing cut off scores for screeners, we allow more false positives since the screener is only supposed to be the first gate of a multiple gate system of identification.

We would also want to make sure that children with certain characteristics (unrelated to their at-risk status) are not being identified more or less often than others as being at-risk for behavioral and emotional maladjustment. For example, if the screener is only identifying those children with externalizing problems and not those with internalizing problem such as anxiety or depression, then those children with internalizing problems would not be identified and would not receive the more comprehensive assessment and services that they need. We would also want to determine whether teachers are over-identifying children of specific gender or ethnicity as well.

Lastly, we would want to evaluate how the children who are identified by the screener are being served. What types of early interventions are in place and are these interventions benefiting children? As Goodman and colleagues (2003) explained, "There would obviously be no point in identifying a greater proportion of child with psychiatric disorders in the community if the only consequence were greater access to ineffective treatments…even if treatments are effective, there is no point identifying more children in need of treatment if existing services are already overstretched and no resources are available" (p. 171).

This type of longitudinal research is costly and time-intensive; however, it is a crucial step in evaluating the effectiveness of a universal screening program. We must not encourage the implementation of large-scale screening programs without extensive evaluation.

Addressing the Public Perceptions of Screening

As demonstrated recently by some high profile legal actions and parental complaints regarding emotional and behavioral screening, public perceptions of screening

must be addressed. One concern is that asking questions about suicidal intent may entice adolescents to actively consider suicide when they would have not done so otherwise. Gould, Marrocco, Kleinman, Thomas, Mostkoff, Cote, and Davies (2005) recently evaluated the iatrogenic risk of youth suicide screening programs and found no evidence to suggest that this is the case. In fact, their findings suggest that the screening may have been beneficial for students with depression symptoms or previous suicide attempts. Although scientific research has not supported the hypothesis of iatrogenic risk of youth suicide screening programs, this public fear represents a significant barrier to public acceptance of universal emotional and behavioral screening of children and adolescents.

In August 2003, Illinois was the first state to pass legislation in which a plan, drafted by the *Illinois Children's Mental Health Partnership* (ICMHP), recommended that "all children receive periodic social and emotional developmental screens" (Barlas, 2004). This plan was met with great opposition by a group of parents who felt that, according to Barbara Shaw, chairman of the (ICMHP), "the schools have no place futzing with their children's mental health." The parents feared that emotional and behavioral screening would lead to the unnecessary labeling and medicating of their children.

These occurrences suggest that public opinion of mental health disorders and our ability to detect and treat them, is not yet at the point where universal emotional and behavioral screening would be widely accepted. However, we can take steps in order to make emotional and behavioral screening as acceptable and "consumer-friendly" as possible. Through research and studying the consequences of universal child mental health screening, we can demonstrate the true effects of universal child mental health screening.

Conclusions

The current study has made important strides in examining the many unanswered questions that still exist with regard to child mental health screening. In summary, we found that the newly-developed BASC-2 screeners appear to be very promising as first gate screening measures. Additionally, parents were found to do a better job than teachers as a first gate screener. Adding a comprehensive behavior rating scale as second gate significantly improved identification accuracy and utilizing a different informant at the second gate generally improved classification accuracy further. Lastly, when implementing a two informant, two gate screening procedure, a screener is a valid option as the second gate in place of a longer behavior rating scale such as the full BASC-2.

Heretofore mental health screening of children has constituted an area of practice based on good intentions and reasonable premises in the absence of a substantial corpus of scientific knowledge. Much more research must still be done in order to ensure that sound science guides the increasingly popular practice of screening children for behavioral and emotional problems that may lead to numerous untoward outcomes.

References

Achenbach, T. M. (1991) Integrative Guide to the 1991 CBCL/4-18, YSR, and TRF Profiles. Burlington, VT: University of Vermont, Department of Psychology

Achenbach, T.M. & Edelbrock, C. (1987). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile.* Burlington: University of Vermont Department of Psychiatry.

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

Altman, D.G. (1991). *Practical Statistics for Medical Research.* London: Chapman and Hall.

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders* (3rd Ed. Revised, DSM-III-R), Washington, DC: Author.

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders* (4th Ed.), Washington, DC: Author.

Aos, S., Lieb, R., Mayfield, J., Miller, M., & Pennucci, A. (2004). *Benefits and costs of prevention and early intervention programs for youth.* Olympia: Washington State Institute for Public Policy.

Arcia, E. & Fernandez, M.C. (2003). Presenting problems and assigned diagnoses among young latino children with disruptive behavior. *Journal of Attention Disorders,* 6, 177-185.

Aronen, E. T., Teerikangas, O. M., & Kurkela, S. A. (1999). The continuity of psychiatric symptoms from adolescence into young adulthood. *Nordic Journal of Psychiatry, 53(5),* 333–338.

August, G. J., Ostrander, R., & Bloomquist, M. J. (1992). Attention Deficit Hyperactivity Disorder: Epidemiological screening method. *American Journal of Orthopsychiatrics, 62,* 387-396.

August, G.J., Realmuto, G.M., Crosby, R.D., & MacDonald, A.W. (1995). Community-based multiple-gate screening of children at risk for conduct disorder. *Journal of Abnormal Child Psychology, 23,* 521-544.

Barlas, S. (2004). Illinois passes controversial child screening plan. *Psychiatric Times.* 21.

Beck, A.T., Kovacs, M., & Weissman, A. (1979). Assessment of suicidal ideation: The Scale of Suicide Ideation. *Journal of Consulting and Clinical Psychology, 47,* 343-352.

Beck, J.S., Beck, A.T., & Jolly, J. (2001). *The Beck Youth Inventories of Emotional and Social Impairment,* San Antonio, TX: Harcourt Assessment Inc.

Becker, A., Woerner, W., Hasselhorn, M., Banaschewski, T., & Rothenberger, A. (2004). Validation of the parent and teacher SDQ in a clinical sample, *European Child and Adolescent Psychiatry, 13,* II/11-II/16.

Bennett, K.J., Lipman, E.L., Brown, S., Racine, Y., Boyle, M.H., & Offord, D.R. (1999). Predicting conduct problems: Can high-risk children be identified in kindergarten and grade 1? *Journal of Consulting and Clinicap Psychology, 67,* 470-480.

Bennett, K.J., & Offord, D.R. (2001). Screening for Conduct Problems: Does the predictive accuracy of conduct disorder symptoms improve with age? *Journal of the American Academy of Child and Adolescent Psychiatry, 40,* 1418-1425.

Biederman, J., Keenan, K., & Faraone, S.V. (1990). Parent-based diagnosis of attention deficit disorder predicts a diagnosis based on teacher report. *Journal of the American Academy of Child and Adolescent Psychiatry, 29,* 698-701.

Boggs, S. R., Eyberg, S., & Reynolds, L. A. (1990). Concurrent validity of the Eyberg Child Behavior Inventory. *Journal of Clinical Child Psychology*, 19 (1), 75-78.

Bonner, M. (2003). [Review of the Beck Youth Inventories of Emotional and Social Impairment]. *Fifteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute.

Borowsky, I.W., Mozayeny, S., &Ireland, M. (2003). Brief psychosocial screening at health supervision and acute care visits. *Pediatrics, 112,* 129-133.

Bose-Deakins, J.A., & Floyd, R.G. (2004). A review of the Beck Youth Inventories of Emotional and Social Impairment. *Journal of School Psychology, 42,* 333-340.

Bourdon, K.H., Goodman, R., Rae, D.S., Simpson, G., & Koretz, D.S. (2005). The strengths and difficulties questionnaire: U.S. normative data and psychometric properties. *Journal of the American Academy of Child and Adolescent Psychiatry, 44,* 557-564.

Boyle, M.H., & Pickles, A.R. (1997). Influence of maternal depressive symptoms on ratings of childhood behavior. *Journal of Abnormal Child Psychology, 25,* 399-412.

Brent, D.A., Baugher, M., Bridge, J., Chen, T., & Chiappetta, L. (1999). Age- and sex-related risk factors for adolescent suicide. *Journal of the American Academy of Child and Adolescent Psychiatry,* 38, 1497- 1505.

Brestan, E.V., & Eyberg, S.M. (1998) Effective psychosocial treatments of conduct-disordered children and adolescents: 29 years, 82 studies, and 5,272 kids. *Journal of Clinical Child Psychology. 27,* 180 – 189.

Burnam, M.A., Wells, K.B., Leake, B., & Landsverk, J. (1988). Development of a brief screening instrument for detecting depressive disorders. *Medical Care, 26,* 775-789.

Burnham, K.P., & Anderson, D.R. (1998). *Model Selection and inference: A practical information-theoretic approach.* New York: Springer-Verlag.

Campaign for Mental Health Reform (2005). A public health crisis: children and adolescents with mental disorders. Congressional briefing. Retrieved September 1, 2005, from www.mhreform.org/kids.

Campbell, J.M., Bell, S.K., & Keith, L.K. (2001). Concurrent validity of the Peabody picture vocabulary test – third edition as an intelligence and achievement screener for low SES African American children. *Assessment, 8,* 85-94.

Carberry, A.T., & Handal, P.J. (1980). The use of the AML scale with a Headstart population: Normative and validation studies. *American Journal of Community Psychology, 8,* 353-363.

Carran, D.T., & Scott, K.G. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education, 12,* 196-211.

Christopher, R. (2001). [Review of the Multidimensional Anxiety Scale for Children]. *Fourteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute.

Clarke-Stewart, K.A., Allhusen, V.D., McDowell, D.J., Thelen, L., & Call, J.D. (2003). Identifying psychological problems in young children: How do mothers compare with child psychiatrists? *Applied Developmental Psychology*, *23,* 589-624.

Cohen, M.A. (1998), The monetary value of saving a high risk youth. *Journal of Quantitative Criminology*, *4,* 5-33

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for Behavioral Sciences (3rd Ed.).* Mahwah, NJ: Erlbaum.

Collett, B.R., Jeneva, L.O., & Myers, K.M. (2003). Ten-year review of rating scales. V: Scales assessing attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42,* 1015-1037.

Comrey, A.L., & Lee, H.B. (1992). *A First Course in Dactor Analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Conners, C. K. (1973). Rating scales for use in drug studies with children. *Psychopharmacology Bulletin*, 9, 24-29.

Conners, C.K., Parker, J.D.A., Sitarenios, G., & Epstein, J.N. (1997). The Revised Conners' Parent Rating Scale (CPRS-R): Factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology*, 26, 257-268.

Cowen, E.L., Dorr, D., Clarfield, S.P., Kreling, B., McWilliams, S.A., Pokracki, F., Pratt, D.M., Terrell, D.L., & Wilson, A.B. (1973). The AML: A quick screening device

for early identification of school maladaption. *American Journal of Community Psychology, 1,* 12-35.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Demaray, M.K., Elting, J., & Schaefer, K. (2003). Assessment of attention-deficit/hyperactivity disorder (ADHD): A comparative evaluation of five, commonly used, published rating scales. *Psychology in the Schools, 40,* 341-361.

Derogatis, L.R. & DellaPietra, L. (1994). Psychological tests in screening for psychiatric disorder. In M.E. Maruish (Ed.), *The Use of Psychological Testing for Treatment Planning and Outcome Assessment* (pp.22 – 54). New Jersey: Lawrence Erlbaum Associates.

Dickey, W.C., & Blumberg, S.J. (2004). Revisiting the factor structure of the strengths and difficulties questionnaire: United States, 2001. *Journal of the American Academy of Child and Adolescent Psychiatry, 43,* 1159-1167.

DuPaul, G.J., Power, T.J., Anastopoulos, A.D., & Reid, R. (1998). *ADHD Rating Scale – IV: Checklists, norms, and clinical interpretation.* New York: The Guilford Press.

Eggert, L., Thompson, E., & Hering, J. (1994). A measure of adolescent potential for suicide (MAPS): development and preliminary findings. *Suicide and Life-Threatening Behavior,* 24, 359 – 381.

Eyberg, S., & Pincus, D. (1999). Eyberg Child Behavior Inventory & Sutter-Eyberg Student Behavior Inventory - Revised. Psychological Assessment Resources, Odessa, FL: Psychological Assessment Rescources.

Eyberg, S.M., & Robinson, E.A. (1983). Conduct problem behavior: Standardization of a behavioral rating scale with adolescents. *Journal of Clinical Child Psychology, 12,* 347-354.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods,* 4, 272-299.

Fantuzzo, J., Bulotsky, R., McDermott, P., Mosca, S., & Lutz, M. N. (2003). A multivariate analysis of emotional and behavioral adjustment and preschool educational outcomes. *School Psychology Review,* 32, 185-203.

Faulstich, M.E., Carey, M.P., Ruggiero, L., Enyart, P., & Gresham, F. (1986). Assessment of depression in childhood and adolescence: An evaluation of the center for epidemiological studies depression scale for children (CES-DC). *American Journal of Psychiatry, 143,* 1024-1026.

Flanagan, K.S., Bierman, K.L., & Kam, C.M. (2003). Identifying at-risk children at school entry: the usefulness of multibehavioral problem profiles. *Journal of Clinical Child and Adolescent Psychology, 32,* 396-407).

Flawes, D.J., & Dadds, M.R. (2004). Australian data and psychometric properties of the strengths and difficulties questionnaire. *Australian and New Zealand Journal of Psychiatry, 38,* 644-651.

Friedman, R.M., Katz-Leavy, J., Manderscheid, R., & Sondheimer, D. (1996). Prevalence of serious emotional disturbance in children and adolescents. In R.W. Manderscheid & M.A. Sonnenschein (Eds.), *Mental Health, United States, 1996* (pp. 71-88). Rockville, MD: Center for Mental Health Services.

Gall, G., Pagano, M.E., Desmond, M.S., Perrin, J.M., & Murphy, J.M. (2000) Utility of psychosocial screening at a school-based health center. *Journal of School Health, 70*, 292-298.

Gardner, W., Murphy, M., Childs, G., Kelleher, K., Pagano, M., Jellinek, M.,McInerny, T. K., Wasserman, R.C., Nutting, P., & Chiappetta, L. (1999) The PSC-17: a brief pediatric symptom checklist including psychosocial problem subscales: a report from PROS and ASPN. *Ambulatory Child Health*, 5, 225 –236.

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

Goodman, R., Ford, T., Corbin, T., & Meltzer, H. (2004). Using the strengths and difficulties questionnaire (SDQ) multi-informant algorithm to screen looked-after children for psychiatric disorders. *European Child and Adolescent Psychiatry, 13,* 11/25-11/31.

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,* 166-172.

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*, 17-24.

Gottlieb, G. (1991). Experiential canalization of behavioral development: Theory. *Developmental Psychology,* 27, 4-13.

Gould, M.S., Marrocco, F.A., Kleinman, M., Thomas, J.G., Mostkoff, K., Cote, J., & Davies, M. (2005). Evaluating iatrogenic risk of youth suicide screening programs. *Journal of the American Medical Association,* 293, 1635-1643.

Goyette, C.H., Conners, C. K., & Ulrich, R.F. (1978). Normative data on revised conners parent and teacher rating scales. *Journal of Abnormal Child Psychology, 6.* 221-236.

Gray, J. (1987). The psychology of fear and stress. New York: Cambridge University Press.

Grietens, H., Onghena, P., Prinzie, P., Gadeyne, E., Van Assche, V., Ghesquiere, P., & Hellinckx, W. (2004). Comparison of mothers', fathers', and teachers' reports on problems behavior in 5- to 6- year-old children. *Journal of Psychopathology and Behavioral Assessment, 26,* 137-146.

Grills, A.E. & Ollendick, T.H. (2003). Multiple informant agreement and the anxiety disorders interview schedule for parents and children. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 30-40.

Grosenick, J.K. (1981). Public school and mental health services to severely behavior disordered students. *Behavior Disorders,* 6, 183-190.

Gutman, L.M., Sameroff, A.J., & Cole, R. (2003). Academic growth curve trajectories from 1st grade to 12th grade: Effects of multiple social risks and preschool child factors. *Developmental Psychology,* 39, 777-790.

Handwerk, M.L., Larzelere, R.E., Soper, S.H., & Friman, P.C. (1999). Parent and child discrepancies in reporting severity of problem behaviors in three out-of-home settings. *Psychological Assessment, 11,* 14-22.

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

Hill, L.G., Lochman, J.E., Coie, J.D., Greenberg, M.T. & The Conduct Problems Prevention Research Group (2004). Effectiveness of early screening for externalizing problems: Issues of screening accuracy and utility. *Journal of Consulting and Clinical Psychology, 72,* 809-820.

Hirshfield-Becker, D.R. & Biederman, J. (2002). Rationale and principles for early intervention with young children at risk for anxiety disorders. *Clinical Child and Family Psychology Review, 5.* 161-172.

Huang, L., Stroul, B., Friedman, R., Mrazek, P., Friesen, B., Pires, S., & Mayberg, S. (2005). Transforming mental health care for children and their families. *American Psychologist,* 60, 615-627.

Jamieson, K.H. & Romer, D. (2005). A call to action on adolescent mental health. In D.L. Evans, E.B. Foa, R.E. Gur, H. Hendin, C. P. O'Brien, M. E. P. Seligman and B. T. Walsh (Eds.), *Treating and Preventing Adolescent Mental Health Disorders: What We Know and What We Don't Know (*pp. 598-615). New York: Oxford University Press.

Jellinek, M. Little, M., Murphy, J.M., & Pagano, M. (1995). The Pediatric Symptom Checklist; support for a role in a managed care environment. *Archives of Pediatrics and Adolescent Medicine,* 149, 740-746.

Jellinek, M. & Murphy, J.M. (1988). Screening for psychosocial disorders in pediatric practice. *American Journal of Diseases of Children,* 109, 371-378.

Jellinek, M. & Murphy, J.M. (1990). The recognition of psychosocial disorders in pediatric office practice: the current status of the pediatric symptom checklist. *Journal of Developmental and Behavioral Pediatrics, 11*, 273-278.

Jellinek, M.S., Murphy, J.M., & Burns, B.J. (1986). Brief psychosocial screening in outpatient pediatric practice. *The Journal of Pediatrics, 109,* 371-377.

Jenkins, R. (1998). Mental health and primary care – implications for policy. *International Review of Psychiatry,* 10, 158-160.

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,* 1569-1579.

Jimerson, S., Egeland, AB., & Teo, A. (1999). A longitudinal study of achievement trajectories: Factors associated with change. *Journal of Educational Psychology,* 91, 116-126.

Johnston, C. & Murray, C. (2003). Incremental validity in the psychological assessment of children and adolescents, *Psychological Assessment,* 15, 496-507.

Jones, D., Dodge, K.A., Foster, E.M., Nix, R., and the Conduct Problems Prevention Research Group. (2002). Early identification of children at risk for costly mental health service use. *Prevention Science, 3,* 247-256.

Kamphaus, R. W., & Frick, P. J. (2002). *Clinical assessment of child and adolescent personality and behavior (2nd ed.).* Boston: Allyn & Bacon.

Kamphaus, R. W., & Reynolds, C. R. (in press). Behavior Assessment System for Children – Second Edition Screening System (BASC-2 SS). Bloomington, MN: Pearson Assessment.

Kamphaus, R.W., Thorpe, J.S., Winsor, A.P., Kroncke, A.P., Dowdy, E.T., & VanDeventer, M.C. (2007). Development and predictive validity of a teacher screener for child behavioral and emotional problems at school. *Educational and Psychological Measurement, 67,* 342-356.

Kaslow, N.J., & Thompson, M.P. (1998). Applying the criteria for empirically supported treatments to studies of psychosocial interventions for child and adolescent depression. *Journal of Clinical Child Psychology, 27,* 146 –155.

Kaufman, J., Cook, A., Arny, L., Jones, B., & Pittinsky, T. (1994). Problems defining resiliency: Illustrations from the study of maltreated children. *Development and Psychopathology, 6,* 215-229.

Kavan, M.G. (1992). [Review of the Children's Depression Inventory]. *Eleventh Mental Measurements Yearbook.* Lincoln, NE: Buros Institute.

Kelleher, K. J., Moore, C. D., Childs, G. E., Angelilli, M. L., & Comer, D. M. (1999). Patient race and ethnicity in primary care management of child behavior problems. *Medical Care, 37,* 1092-1104.

Kessler, R.C., Stang, P., Wittchen, H.U., Stein, M., & Walters, E.E. (1999). Lifetime comorbidities between social phobia and mood disorders in the US National Comorbidity Survey. *Psychological Medicine, 29,* 555-567.

Kim, S, & Rowe, E. (2004, April). Similar risks, dissimilar outcomes: Rethinking risk in educational context. Paper presented at the annual conference of the American Educational Research Association, San Diego, CA.

Knoff, H.M. (1992). [Review of the Children's Depression Inventory]. *Eleventh Mental Measurements Yearbook.* Lincoln, NE: Buros Institute.

Knoff, H.M. (2001). [Review of the Conners' Ratings Scales-Revised]. *Fourteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute.

Kolko, D.J., & Kazdin, A.E. (1993). Emotional/behavioral problems in clinic and nonclinic children: correspondence among children, parent, and teacher reports. *Journal of Child Psychology and Psychiatry, 34,* 991-1006.

Kobak, K.A., Taylor, L.vH., Dottl, S.L., Greist, J.H., Jefferson, J.W., Burroughs, D., Katzelnick, D.J., & Mandell, M. (1997). Computerized screening for psychiatric disorders in an outpatient community mental health clinic. *Psychiatric Services,* 48, 1048-1057.

Kovacs, M. (1992). *Children's Depression Inventory (CDI) manual.* New York: Multi-Health Systems.

Kresanov, K., Tuominen, J., Piha, J., & Almqvist, F. (1998). Validity of child psychiatric screening methods. *European Child and Adolescent Psychiatry, 7,* 85-95.

Lett, N.J. & Kamphaus, R.W. (1997). Differential validity of the BASC Student Observation System and BASC Teacher Rating Scale. *Canadian Journal of School Psychology,* 13, 1-14.

Leon, A.C., Kathol, R., Portera, L., Farber, L., Olfson, M., Lowell, K.N., & Sheehan, D.V. (1999). Diagnostic errors of primary care screens for depression and panic disorder. *International Journal of Psychiatry in Medicine,* 29, 1-11.

Leung, P.W.L., Lucas, C.P., Hung, S., Kwong, S., Tang, C., Lee, C., Ho, T., Lieh-Mak, F., & Shaffer, D. (2005). The test-retest reliability and screening efficiency of DISC predictive scales-version 4.32 (DPS-4.32) with Chinese children/youths. *European Child and Adolescent Psychiatry, 14,* 461-465.

Littrup, P.J., Lee, F., & Mettlin, C. (1992). Prostate cancer screening: Current trends and future implications. *A Cancer Journal for Clinicians, 42,* 198-211.

Lloyd, J.W., Kauffman, J.M., Landrum, T.J., & Roe, D.L. (1991). Why do teachers refer pupils for special education: An analysis of referral records. *Exceptionality,* 2, 115-126.

Lochman, J. E. and the Conduct Problems Prevention Research Group (1995). Screening of child behavior problems for prevention programs at school entry. *Journal of Consulting and Clinical Psychology, 63*, 549-559.

Loeber, R., Dishion, T.J., & Patterson, G.R. (1984). Multiple gating: A multistage assessment procedure for identifying youths at risk for delinquency. *Journal of Research on Crime and Delinquency, 21,* 7-32.

Loeber, R., Green, S.M., & Lahey, B.B. (1990). Mental health professionals' perceptions of the utility of children, mothers, and teachers as informants on childhood psychopathology. *Journal of Clinical Child Psychology, 19,* 136-143.

Loeber, R., Green, S.M., Lahey, B.B., Stouthamer-Loeber, M. (1991). Differences and similarities between children, mothers, and teachers as informants on disruptive child behavior. *Journal of Abnormal Child Psychology, 19,* 75 – 95.

Lucas, C.P., Zhang, H., Fisher, P.W., Shaffer, D., Regier, D.A., Narrow, W.E., Bourdon, K., Dulcan, M.K., Canino, G., Rubio-Stipec, M., Lahey, B.B., & Friman, P. (2001). The DISC Predictive Scales (DPS): efficiently screening for diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry,* 40, 443-449.

Malmberg, M., Rydell, A.M., & Smedje, H. (2003). Validity of the Swedish version of the strengths and difficulties questionnaire (SDQ-Swe). *Nordic Journal of Psychiatry, 57,* 357-364.

March, J.S., Parker, J.D., Sullivan, K., Stallings, P., & Conners, C.K. (1997). The Multidimensional Anxiety Scale for Children (MASC): factor structure, reliability, and validity. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 554-565.

March L.S., & Sullivan, K. (1999). Test-retest reliability of the multidimensional anxiety scale for children. *Journal of Anxiety Disorders, 13,* 349-358.

Martin, Sarah (2005). Healthy kids make better kids. *Monitor on Psychology,* 36, 24-26.

Masten, A.S. & Coatsworth, J.D. (1998). The development of competence in favorable and unfavorable environments: Lessons from research on successful children. *American Psychologist,* 53, 205-220.

Matthey, S., & Petrovski, P. (2002). The Children's Depression Inventory: Error in cutoff scores for screening purposes. *Psychological Assessment, 14*, 146-149.

McCarney, S. & Leigh, J. (1990). *McCarney Behavior Evaluation Scale – 2.* Columbia, MO: Educational Services.

McDermott, P., Marston, N. & Stott, D. (1994). *McDermott Adjustment Scales for Children and Adolescents.* Phoenix, AZ: Ed. And Psych Associates.

McEvoy, A., & Welker, R. (2000). Antisocial behavior, academic failure, and school climate: a critical review. *Journal of Emotional and Behavioral Disorders,* 8, 130-141.

McFall, R.M. (2005). Theory and utility – Key themes in evidence-based assessment: Comment on the special section. *Psychological Assessment*, 17, 312-323.

McGoey, K.E., Eckert, T.L., & Dupaul, G.J. (2002). Early intervention for preschool-age children with ADHD: A literature review. *Journal of Emotional and Behavioral Disorders, 10,* 1-27.

McGrew, K. S., & Woodcock, R. W. (2001). Technical manual. *Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.

Meehl, P.E., & Rosen, A.(1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194-216.

Meikamp, J. (2003). [Review of the Eyberg Child Behavior Inventory and Sutter-Eyberg Student Behavior Inventory-Revised]. *Fifteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute.

Mellor, D. (2004). Furthering the use of strengths and difficulties questionnaire: Reliability with younger child respondents. *Psychological Assessment, 16,* 396-401.

Mesman, J. & Koot, H.M. (2000). Child-reported depression and anxiety in preadolescence: I. Associations with parent- and teacher- reported problems. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 1371-1378.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Murphy J.M., Jellinek, M.S., & Milinsky, S. (1989). The pediatric symptom checklist: validation in the real world of middle school. *Journal of Pediatric Psychology,* 14, 629-639.

Murphy, J.M., Reede, J., Jellinek, M.S., & Bishop, S.J. (1992). Screening for psychosocial dysfunction in inner-city children: further validation of the Pediatric Symptom Checklist. *Journal of the American Academy of Child and Adolescent Psychiatry,* 31, 1105-1111.

National Mental Health Association (2005). Study shows mental illness often begins in youth, treatment delays worsen issues. *The Bell: The newsletter of the National Mental Health Association,* July 2005, 1-5.

New Freedom Commission on Mental Health (2003). *Achieving the Promise: Transforming Mental Health Care in America, Final report.* DHHS Pub. No. SMA – 0303832. Rockville, MD. Retrieved September 1, 2005 from http://www.mentalhealthcommission.gov/.

Newman, D.L., Moffitt, T., Caspi, A., & Silva, P.A. (1998). Comorbid mental disorders: Implications for treatment and sample selection. *Journal of Abnormal Psychology, 107,* 305-311.

Noel, M. (1982). Public school programs for the emotionally disturbed: Overview. In M. Noel & N. Haring (Eds.), *Progress or change: Issues in educating the emotionally disturbed (*Vol. 2, pp. 1-28). Seattle: University of Washington.

Olafsen, K. S., & Sommerfelt, K. (1999). The Yale Children's Inventory - A screening tool for attention deficits and related disorders: Normative data for boys. *Scandinavian Journal of Psychology, 40,* 121-125.

Pagano, M. E., Cassidy, L. J., Little, M., Murphy, J. M., & Jellinek, M. S. (2000). Identifying psychosocial dysfunction in school aged children: The Pediatric Symptom Checklist as a self-report measure. *Psychology in the Schools, 37,* 91-106.

Pedhazur, E.J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction (3rd Ed.).* Fort Worth, TX: Harcourt Brace College Publishers.

Pelham, W.E. Jr.,Wheeler, T., & Chronis, A. (1998) Empirically supported psychosocial treatments for attention deficit hyperactivity disorder. *Journal of Clinical Child Psychology, 27,* 190 –205.

Petras, H., Chilcoat, H.D., Leaf, P.J., Ialongo, N.S., & Kellam, S.G. (2005). Utility of TOCA-R scores during the elementary school years in identifying later violence among adolescent males. *Journal of the American Academy of Child and Adolescent Psychiatry, 43,* 88-96.

Petras, H., Ialongo, N., Lambert, S.F., Barreuco, S., Schaeffer, C.M., Chilcoat, H., & Kellam, S. (2004). The utility of elementary school TOCA-R scores in identifying later criminal court violence among adolescent females. *Journal of the American Academy of Child and Adolescent Psychiatry, 44,* 790-797.

Phillips, V., Nelson, C.M., & McLaughlin, J.R. (1993). Systems change and services for students with emotional/behavioral disabilities in Kentucky. *Journal of Emotional and Behavioral Disorders,* 1, 155-164.

Power, T.J., Andrews, T.J., Eiraldi, R.B., Doherty, B.J., Ikeda, M.J., DuPaul, G.J, & Landau, S. (1998). Evaluating attention deficit hyperactivity disorder using multiple informants: the incremental utility of combining teacher with parent reports. *Psychological Assessment, 10,* 250-260.

Rapport, M.D., Denney, C.B., Chung, K.M., & Hustace, K. (2001). Internalizing behavior problems and scholastic achievement in children: Cognitive and behavioral pathways as mediators of outcome. *Journal of Clinical Child Psychology,* 30, 536-551.

Rettew, D.C., Copeland, W., Stanger, C., & Hudziak, J.J. (2004). Associations between temperament and DSM-IV externalizing disorders in children and adolescents. *Developmental and Behavioral Pediatrics, 25,* 383-391.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children-Second Edition (BASC-2).* Circle Pines, MN: AGS.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior Assessment System for Children (BASC).* Circle Pines, MN: AGS.

Reynolds, C.R. & Richmond, B.O. (1985). *"What I think and Feel" Reynolds Child Manifest Anxiety Scale (RCMAS).* Los Angeles, CA: Western Psychological Services.

Reynolds, W.M. (1988). *Suicidal Ideation Questionnaire (SIQ): Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Reynolds W.M. (1989). *Reynolds Child Depression Scale,* Lutz, FL: Psychological Assessment Resources, Inc.

Reynolds, W.M. (1991). A school-based procedure for the identification of adolescents at risk for suicidal behaviors. *Family and Community Health*, 14, 64-75.

Reich, W., & Welner, Z. (1990). *The Diagnostic Interview for Children and Adolescents – Revised (DICA –R).* Structured psychiatric interview. St. Louis, MO: Washington University.

Renk, K. (2005). Cross-informant ratings of the behavior of children and adolescents: The "Gold Standard." *Journal of Child and Family Studies, 14,* 457-468.

Ringel, J. & Sturm, R. (2001). National estimates of mental health utilization and expenditure for children in 1998. *Journal of Behavioral Health Services and Research,* 28, 319-332.

Romer, D. & McIntosh, M. (2005). The roles and perspectives of school mental health professionals in promoting adolescent mental health. In D.L. Evans, E.B. Foa, R.E. Gur, H. Hendin, C. P. O'Brien, M. E. P. Seligman and B. T. Walsh (Eds.), *Treating and Preventing Adolescent Mental Health Disorders: What We Know and What We Don't Know (*pp. 598-615). New York: Oxford University Press.

Ronning, J.A., Handegaard, B.H., Sourander, A., Morch, W.T. (2004). The strengths and difficulties self-report questionnaire as a screening instrument in Norwegian community samples. *European Child and Adolescent Psychiatry, 13*, 73-82.

Rothbart, M.K., & Bates, J.E. (1998). Temperament. In Damon, W. (Series Ed.) and Eisenberg, N. (Ed.), *Handbook of Child Psychology: Vol. 3, Social, emotional, and personality development* (5th ed., pp. 105-176). New York: Wiley.

Rutter, M., & Sroufe, L.A. (2000). Developmental psychopathology: Concepts and challenges. *Development and Psychopathology, 12*, 265-296.

Rynn, M.A., Barber, J.P., Khalid-Khan, S., Siqueland, L., Dembiski, M., McCarthy, K.S., & Gallop, R. (2006). The psychometric properties of the MASC in a pediatric psychiatric sample. *Anxiety Disorders, 20*, 139-157.

Saunders, S.M. & Wojcik, J.V. (2004). The reliability and validity of a brief self-report questionnaire to screen for mental health problems: the health dynamics inventory. *Journal of Clinical Psychology in Medical Settings,* 11, 233 – 241.

Schmidt, N.B., Kotov, R., & Joiner, T.E. (2004). Taxometrics: toward a new diagnostic scheme for psychopathology. Washington, DC: American Psychological Association.

Schmitz, N., Kruse, J., Heckrath, C., Alberti, L., & Tress, W. (1999). Diagnosing mental disorders in primary care: the General Health Questionnaire (GHQ) and the Symptom Check List (SCL-90-R) as screening instruments. *Social Psychiatry & Psychiatric Epidemiology, 34*, 360-367.

Schwab-Stone, M., Shaffer, D., Dulcan, M., Jensen., P., Fisher, P., Bird, H., Goodman, S., Lahey, B., Lichtman, J., Canino, G., Rubio-Stipec, M., & Rae, D. (1996). Criterion validity of the NIMH Diagnostic Interview Schedule for Children Version 2.3 (DISC-2.3). *Journal of the American Academy of Child and Adolescent Psychiatry,* 35, 878-888.

Seligman, L.D., & Ollendick, T.H. (1998). Comorbidity of anxiety and depression in children and adolescents: An integrative review. *Clinical Child and Family Psychology Review, 1,* 125-144.

Seligman, L.D., Ollendick, T.H., Langley, A.K., & Baldacci, H.B. (2004). The utility of measures of child and adolescent anxiety: a meta-analytic review of the revised children's manifest anxiety scale, the state-trait anxiety inventory for children, and the child behavior checklist. *Journal of Clinical Child and Adolescent Psychology, 33,* 557-565.

Shaffer, D., Fisher, P., Lucas, C., Dulcan, M., & Schwab-Stone, M. (2000). NIMH Diagnostic Interview Scale for Children version IV: description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry,* 39, 28-38

Shaffer, D., Scott, M., Wilcox, H., Maslow, C., Hicks, R., Lucas, C.P., Garfinkel, R., & Greenwald, S. (2004). The Columbia SuicideScreen: Validity and reliability of a screen for youth suicide and depression. *Journal of the American Academy of Child and Adolescent Psychiatry,* 41, 71-79.

Shaywitz, S.E., Schnell, C., Shaywitz, B.A., & Towle, V.R. (1986). Yale Children's Inventory (YCI): an instrument to assess children with attentional deficits and

learning disabilities I. scale development and psychometric properties. *Journal of Abnormal Child Psychology,* 14, 347 – 364.

Shaywitz, S.E., Shaywitz, B.A., Schnell, C., & Towle, V.R. (1988). Concurrent and predictive validity of the Yale Children's Inventory: An instrument to assess children with attentional deficits and learning disabilities. *Pediatrics, 81,* 562-571.

Shea, M. Widiger, T., & Klein, M. (1992). Comorbidity of personality disorders and depression: Implications for treatment. *Journal of Clinical and Consulting Psychology, 60,* 857-868.

Sickles, E.A., Ominsky, S.H., Sollitto, R.A., Gavlin, H.B., & Monticciolo, D.L. (1990). Medical audit of a rapid-throughout mammography screening practice: Methodology and results of 27,114 examinations. *Radiology, 175,* 323-327.

Simonian, S.J. & Tarnowski, K.J. (2001). Utility of the pediatric symptom checklist for behavioral screening of disadvantaged children. *Child Psychiatry and Human Development,* 31, 269 – 278.

Smith, J.V. (2001). [Review of the AD/HD Comprehensive Teacher's Rating Scale, Second Edition]. *Fourteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute.

Soost, H.J., Lange, H.J., Lehmacher, W., & Ruffing-Kullman, B. (1991). The validation of cervical cystology: Sensitivity, specificity, and predictive values. *Acta Cytologica, 35,* 8-14.

Sourander, A., Helstela, L., & Helenius, H. (1999). Parent-adolescent agreement on emotional and behavioral problems. *Social Psychiatry Psychiaticr Epidemiology, 34,* 657-663.

Spielberger, C.D. (1973). *State Trait Anxiety Inventory for Children.* Palo Alto, CA: Consulting Psychological Press.

Steer, R.A., Kumar, G., Beck, A.T., & Beck, J.S. (2005). Dimensionality of the beck youth inventories with child psychiatric outpatients. *Journal of Psychopathology and Behavioral Assessment, 27,* 123-131.

Stoppelbein, L., Greening, L., Jordan, S.S., Elkin, T.D., Moll, G., & Pullen, J. (2005). Factor analysis of the pediatric symptom checklist with a chronically ill pediatric population. *Developmental and Behavioral Pediatrics*, *26,* 349-355.

Swanson, J., & Carlson, C. L. (1994). DSM-IV rating scale for ADHD and ODD. Unpublished manuscript.

Tarnopolsky, A., Hand, D.J., McLean, E.K., Roberts, H., & Wiggins, R.D. (1979). Validity and uses of a screening questionnaire (GHQ) in the community. *British Journal of Psychiatry,* 134, 508 – 515.

Thompson. E.A. & Eggert, L.L. (1999). Using the Suicide Risk Screen to identify suicidal adolescents among potential high school dropouts. *Journal of the American Academy of Child and Adolescent Psychiatry,* 38, 1506-1514.

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,* 149-157.

Tolan, P.H. & Dodge, K.A. (2005). Children's mental health as a primary care and concern: A system for comprehensive support and service. *American Psychologist,* 60, 601-614.

Ullmann, R.K., Sleator, E.K., & Sprague, R.L. (1988). *ADD-H Comprehensive Teacher's rating Scale (ACTERS).* Champaign, IL: MetriTech, Inc.

Ullmann, R.K., Sleator, E.K., & Sprague, R.L. (2000). *ACTeRS teacher and parent forms manual.* Champagne, IL: Metritech, Inc.

United States Department of Health and Human Services (1999). *Mental Health: A Report of the Surgeon General.* Rockville, MD: U.S. Department of Health and Human Services, Substance abuse, and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health. Retrieved September 1, 2005, from http://www.surgeongeneral.gov/library/mentalhealth/home.html.

United States Public Health Service (2000). *Report of the Surgeon General's Conference on Children's Mental Health: A National Action Agenda,* Washington D.C.: Department of Health and Human Services. Retrieved September 1, 2005, from http://www.surgeongeneral.gov/topics/cmh/childreport.htm.

van der Ende, J. (1999). Multiple informants: multiple views. In H.M. Koot, A.A.M. Crijnen, & R.F. Ferdinand (Eds.), *Child Psychiatric Epidemiology. Accomplishments and Future Directions* (pp. 39-52). Assen, the Netherlands: Van Gorcum.

Van Widenfelt, B.M., Goedhart, A.W., Treffers, P.D., & Goodman, R. (2003). Dutch version of the strengths and difficulties questionnaire (SDQ). *European Child and Adolescent Psychiatry, 12,* 281-289.

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,* 75-81.

Vostanis, P. (2006). Strengths and difficulties questionnaire: research and clinical applications. *Current Opinion in Psychiatry, 19,* 367-372.

Walker, W.O., LaGrone, R.G., & Atkinson, A.W. (1989). Psychosocial screening in pediatric practice: identifying high-risk children. *Journal of Developmental and Behavioral Pediatrics,* 10, 134-138.

Walker, H. M., & McConnell, S. R. (1995). *Walker-McConnell Scale of Social Competence and School Adjustment, Elementary Version: User's Manual*. San Diego, CA: Singular Publishing Group, Inc.

Walker, H., & Severson, H. (1990). *Systematic screening for behavior disorders (SSBD).* Long-mont, CO: Sopris West.

Walker, H., & Severson, H. (1994). Replication of the Systematic Screening for Behavior Disorders (SSBD) procedure for the identification of at-risk children. *Journal of Emotional and Behavioral Disorders,* 2, 66- 78.

Walker, H.M., Severson, H., Stiller, B., Williams, G., Haring, N., Shinn, M., & Todis, B. (1988). Systematic screening of pupils in the elementary-age range at risk for behavior disorders: Development and trial testing of a multiple-gating model. *Remedial and Special Education,* 9, 8-14.

Weis,R., Lovejoy, M.C., & Lundahl, B.W. (2005). Factor structure and discriminative validity of the Eyberg child behavior inventory with young children. *Journal of Psychopathology and Behavioral Assessment, 27,* 269-278.

Weisz, J. R., Sandler, I. N., Durlak, J. A. & Anton, B. S. (2005). Promoting and protecting youth mental health through evidence-based prevention and treatment. *American Psychologist,* 60, 628-648.

Werthamer – Larsson, L., Kellam, S.G., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology, 19,* 585-602.

Whiston, S.C. & Bouwkamp, J.C. (2003). [Review of the Eyberg Child Behavior Inventory and Sutter-Eyberg Student Behavior Inventory-Revised]. *Fifteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute.

Wilson J.M. & Jungner, F. (1968). *Principles and practices of screening for diseases.* Geneva: WHO.

Winter, L. B., Steer, R. A., Jones-Hicks, L., & Beck, A. T. (1999). Screening for major depression disorders in adolescent medical outpatients with the Beck Depression Inventory for primary care. *Journal of Adolescent Health, 24,* 389-394.

Wolraich, M.L., Lambert, E.W., Bickman, L., Simmons, T., Doffing, M.A., & Worley, K.A. (2004). Assessing the impact of parent and teacher agreement on diagnosing attention-deficit hyperactivity disorder. *Developmental and Beahvioral Pediatrics, 25,* 41-47.

World Health Organization (1993). *The ICD-10 Classification of mental and behavioural disorders: Diagnostic criteria for research. World Health Organization.* Geneva, Switzerland: Author.

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology, 68,* 1038-1050.

Zimmerman, M., McDermut, W., & Mattia, J. (2000). Frequency of anxiety disorders in psychiatric outpatients with major depressive disorder. *American Journal of Psychiatry, 157,* 1337-1340.

Table 1. BASC TRS Screener Prediction of Behavioral, Emotional, and Academic Outcomes in Follow up Year: Zero-order Partial Correlations (Kamphaus, et al., 2007)

Outcomes	BASC TRS-C Screener (Yr 2)	BASC Behavioral Symptoms Index (BSI) (Yr 2)
BASC Depression Scale	.370	.401
BASC Anxiety Scale	.195	.202
BASC Withdrawal Scale	.352	.311
BASC Atypicality Scale	.479	.496
BASC Conduct Problems Scale	.497	.437
BASC Social Skills scale	-.471	-.354
Special Education Placement	.306	.287
Pre-referral intervention	.308	.262
Reading Grades	-.546	-.424
Math Grades	-.477	-.355
Work Habits Grades	-.434	-.352
Standardized Reading Score	-.575	-.440
Standardized Math Score	-.547	-.431

* all have p=.000

Table 2. Summary of Available Screening Instruments

Instrument Type	Instrument	Conditions Addressed	Informants and Age Ranges	# of Items	Norming	*Reliability/Validity
Broad	Pediatric Symptom Checklist (PSC; Jellinek, Murphy, & Burns, 1986),	Psychosocial Risk	Parent (6-16) Self-report (11-16)	35 items	206 children, ages 6-12 from three pediatrician's offices – 99% Caucasian, SES (18% high, 44% middle, 38% low); clinical sample of 31 6-12 year olds, all caucasian	Adequate reliability and validity; Feasible in school settings; No PPV or NPV info provided; Parent form: sensitivity from .77 to .95 and specificity from .68 to 1.0; Self-report: sensitivity of .94, specificity of .88 (Jellinek, Little, Murphy, & Pagano, 1995; Simonian & Tarnowski, 2001; Murphy, Jellinek, & Milinsky, 1989; Borowsky, Mozayeny, & Ireland, 2003)
Broad	Pediatric Symptom Checklist - 17 (PSC-17; Gardner, Murphy, Childs, Kelleher, et al., 1999)	Psychosocial Risk	Parent (4-16)	17 items	406 children, ages 4-15, recruited from outpatient/inpatient programs, school-based clinics, and physicians; 71% male	Adequate reliability and validity; Adequate sensitivity and specificity at .82 and .81 respectively; Low PPV of .15 (Gardner et al., 1999; Gardner et al., 2004; Borowsky et al., 2003); More external studies needed
Broad	Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997)	Conduct Problems, Inattention-Hyperactivity, Emotional Symptoms, Peer Problems, and Prosocial Behavior as well as a Total Difficulties Score	Parent (4-16), Teacher (4-16), and Self-report (11-16)	25 items	Originally created in Great Britain, but an American English version has been developed and tested on 9, 577 children in a US national population sample	Adequate overall reliability and validity; reliability of specific scales is questionable; British version found sensitivity of .633, specificity of .946. PPV of .527, NPV of .964(Goodman & Scott, 1999; Mellor, 2004; Goodman et al., 2003); need more research on US version
Specific	Beck Youth Inventories of	depression, anxiety, anger,	Self-report (7-14)	Five 20-item screens	800 children (7-14) stratified based on	Adequate reliability and convergent validity; however

	Emotional and Social Impairment (BDI; Beck, Beck, & Jolly, 2001)	disruptive behavior, and self-concept			1999 Census data on, sex, SES and ethnicity; no indication of stratification by geography	discriminant validity evidence is questionable - majority of scales seem to measure same construct (Bose-Deakins & Floyd, 2004)
Specific	DISC (Diagnostic Interview Schedule for Children) Predictive Scales (DPS – 4.32; Leung et al., 2005)	18 DSM disorders	Self-report (8-18) Parent (8-18)	18 scales (total of 98 items) 14 scales (total of 92 items)	Original: 1,286 subjects, ages 9-17, from 4 sites (Atlanta, New Haven, New York, and Puerto Rico)	Adequate reliability and validity; Sensitivity of .68, specificity of .91, PPV of .34, NPV of .98 (Lucas et al., 2001; Leung et al., 2005); need more external validity studies
Specific	Yale Children's Inventory (YCI; Shaywitz, Schnell, Shaywitz, & Towle, 1986)	Learning disabilities with emphasis on attention deficits	Parent (8-14)	40 items	260 children ages 8 - 14 from CT; SES information provided	Adequate reliability and validity with sensitivity of .875 and specificity of .94 (Olafsen & Sommerfelt, 1999)
Specific	Conners Rating Scales – Revised (CRS-R; Conners, 1973; Conners, Parker, Sitarenios, & Epstein,1998)	Oppositional, Cognitive Problems, Hyperactivity	Parent (3-17) and Teacher (3-17) and self-report (12-17)	Short form(parent: 27 items; teacher: 28 items; self: 27 items) ADHD Index (12 items) DSM-IV Symptom Checklist (18 items)	8000+ sample, ages 3-17 from over 200 sites nationally; manual says ethnicity, age, gender adequately represented, but no specifics given	Adequate reliability and validity; Sensitivity from .78 to .92, specificity from .84 to .94, PPV from .83 to .94, and NPV from .81 to .92 (Connors, et al., 1998); Feasible in school settings; Questionable cut-off scores
Specific	AD/HD Comprehensive Teacher's rating scale – 2nd edition (ACTeRS; Ullman, Sleator,	ADHD	Teacher (K – 8th grade) Parent (K- 8th grade)	24 items 25 items	2, 362 students (k-8), no demographic information provided; separate gender norms available	Insufficient reliability evidence; Evidence of discrimination between ADHD and controls, manual lacks validity evidence

	& Sprague, 1988)					
Specific	ADHD Rating Scale-IV (ADHD-IV, DuPaul et al., 1998)	ADHD	Teacher (5-18) Parent (5-18)	18 items 18 items	National sample of 2,000 children ages 4-20 matched to 1990 U.S. Census data	Excellent reliability and validity; however, parent form has low specificity evidence; Parent form: sensitivity from .83 to .84, specificity low at .49, PPV from .54 to .58, and NPV from .77 to .81; Teacher form: sensitivity from .63 to .72. specificity at .86, PPV from .78-.79, and NPV from .73 to .81 (DuPaul et al., 1998)
Specific	Eyberg Child Behavior Inventory; (Eyberg & Pincus, 1999)	Conduct Problems	Parent (2-16)	36 items	Restandardized in 1999 with 798 children representative of the population in southeastern US on gender, age, ethnicity, SES	Adequate reliability and validity; Sensitivity from .63 to .96, specificity from .87 to >.90, PPV from .63 to .88 and NPV from .82 to .96 (Eyberg & Robinson, 1983; Boggs, Eyberg, & Reynolds, 1990; Rich & Eyberg, 2001; Weise, Lovejoy, & Lundahl, 2005)
Specific	Sutter-Eyberg Student Behavior Inventory – Revised (Eyberg & Pincus, 1999)	Conduct Problems	Teacher (2-16)	38 items	Problematic norms; 415 elementary school children from 11 schools in Gainesville, FL rated by 52 teachers	Some preliminary reliability and validity evidence; No reliability or validity evidence for older children; Need further research
Specific	Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds & Richmond, 1985)	Anxiety	Self-report (6-19)	37 items	4, 972 children from 13 states; Gender, age, and ethinicity given; lack Hispanic sample and no stratification on SES	Adequate composite reliability – subscales lower (in the .6- .7 range); Good criterion-related validity, but unable to differentiate between disorders (Kamphaus & Frick, 2002)
Specific	State-Trait Anxiety Inventory for Children (Spielberger, 1973)	Anxiety	Self-report (9-12)	Two 20 item scales	737 male, 814 female 4th, 5th, and 6th grade elementary school children from six different schools; normative info provided in manual	Self-report: adequate reliability and validity (Carey, Faulstich, & Carey, 1994; Southam-Gerow, Flannery-Schroeder, & Kendall, 2003); Unable to differentiate between disorders

Specific	Multidimensional Anxiety Scale for Children (March, Parker, Sullivan, Stallings, & Conner, 1997)	Anxiety	Self-report (8-19)	39 items (short form: 10)	2, 698 children; lacking hispanic representation (.7%)	Adequate composite reliability – some subscales lower, and initial validity (March, Parker, Sullivan, Stallings, & Connors, 1997); more studies needed to assess discriminant validity; short form has low reliability and lacks validity evidence
Specific	Reynolds Child Depression Scale (RCDS; Reynolds, 1989)	Depression	Self-report (8-12)	29 items	1,620 children from Midwest and CA; SES not controlled, lacking census data	Strong reliability; Good content, criterion-related validity, specificity of 97% and sensitivity of 73%, but highly correlated with anxiety measures (Kamphaus & Frick, 2002)
Specific	Children's Depression Inventory (CDI; Kovacs, 1992)	Depression, negative affect	Parent (8- 17) Teacher (8-17) Self-report (8-17)	Parent: 17 items Teacher:12 items Self-report: 27 items Self-report Short Form: 10 items	1463 children; local norms; little evidence of national representation	Mixed reliability and validity findings; Sensitivity from .729 to .906, specificity from .142 to .94, PPV from .136 to .63, and NPV from .909 to .98 (Matthey & Petrovski, 2002);
Specific	Center for Epidemiological Studies Depression Scale Modified for Children (CES-DC; Faulstich, Carey, Ruggiero, Enyart, & Gresham,1986)	Depression	Self-report (6-17)	20 items	Adapted from adult CES-D which was validated on three samples in Kansas City, MO (n=1173) Washington County, MD (n1=1673, n2=1089) using household interview surveys.	Poor reliability and validity in children, but better for adolescents (Faulstich, Carey et al., 1986); More research is needed
Specific	Columbia Depression Scale (CDS;Shaffer et al., 2000)	Depression, suicide	Self-report (11-17)	22 items	Derived by selecting items from the DISC – no norming sample	Lacking reliability and validity evidence (Shaffer et al., 2000)

Specific	Beck Scale for Suicidal Ideation (Beck, Kovacs, & Weissman, 1979)	Suicidal Risk	Self-report (adolescents and adults)	21 items	178 adults in psychiatric outpatient and inpatient settings – insufficient adolescent sample	No reliability and validity information available for adolescents (Beck, Kovacs, & Weissman, 1979)
Specific	Suicidal Ideation Questionnaire Jr. (SIQ-JR), Suicidal Ideation Questionnaire (SIQ), (Reynolds, 1988, 1991)	Suicidal Risk	Self-report (7-9 grade - SIQ JR; 10-12 grade - SIQ)	SIQ Jr – 15 items SIQ – 30 items	Convenience 7-9 grade sample of 1,290 for SIQ-JR; convenience 10-12 grade sample of 890 for SIQ; from three Midwestern schools; info on gender representation is provided	Adequate reliability; Adequate convergent validity; Good sensitivity ranging from .83 to 1.0; Low specificity from 40% to 70%; questionable cut score (Reynolds, 1991)
Specific	The Suicide Risk Screen (Eggert et al., 1994),	suicide ideation, suicide attempts, depression, and substance use	Self-report (14 years and older)	20 items – embedded in the Health Status Questionnaire 2.0	Norms not provided	Adequate reliability and validity; Good sensitivity at .87 to 1.0, but lower specificity ranging from .54 to .64 (Thompson & Eggert, 1999)
Specific	Columbia Health/ Suicide Screen (CSS; Shaffer et al., 2004)	Depression, suicide	Self-report (11-18)	14 items	Convenience sample of 1,729 9^{th}-12^{th} graders from 7 NY high schools	PPV is low at 12-16%; NPV 99%; Sensitivity of .75 to .88 and specificity of .83 (Shaffer et al., 2004); more research needed

*For screening purposes only –not as diagnostic, but rather an indication for further assessment

Table 3. Teacher (TRS) screener – Child and Adolescent version

Item #	Scale	Item
1	Atn	Pays attention. (r)
2	Cnd	Disobeys.
3	Dep	Is sad.
4	Cnd	Breaks the rules.
5	Std	Is well organized. (r)
6	Hyp	Has poor self-control.
7	Dep	Is easily upset.
8	Lrn	Completes assignments incorrectly because of not following directions.
9	Led	Is good at getting people to work together. (r)
10	Lrn	Has trouble keeping up in class.
11	Anx	Worries about things that cannot be changed.
12	Dep	Says, “Nobody likes me.”
13	Agg	Annoys others on purpose
14	Anx	Is fearful.
15	Som	Has headaches.
16	Atn	Is easily distracted from class work.
17	Fun	Is effective when presenting information to a group. (r)
18	Cnd	Gets into trouble.
19	Led	Gives good suggestions for solving problems. (r)
20	Dep	Is negative about things.
21	Hyp	Disrupts other children’s activities.
22	Som	Complains about health.
23	Atn	Has trouble concentrating.
24	Std	Has good study habits. (r)
25	Anx	Worries.
26	Atn	Has a short attention span.
27	Skl	Encourages others to do their best. (r)

* r = reversed item

*Dep = Depression, Anx = Anxiety, Som = Somatization, Hyp = Hyperactivity, Cnd = Conduct Problems, Agg = Aggression, Atn = Attention Problems, Lrn = Learning Problems, Fun = Functional Communication, Led = Leadership, Skl = Social Skills, Std = Study Skills

Table 4. Parent (PRS) screener – Child and Adolescent version

Item #	Scale	Item
1	Atn	Pays attention. (r)
2	Cnd	Disobeys.
3	Fun	Tracks down information when needed. (r)
4	Cnd	Breaks the rules.
5	Skl	Tries to bring out the best in other people. (r)
6	Hyp	Acts out of control.
7	Dep	Seems lonely.
8	Atn	Is easily distracted.
9	Led	Is good at getting people to work together. (r)
10	Agg	Defies people in authority.
11	Anx	Worries about things that cannot be changed.
12	Adl	Acts in a safe manner. (r)
13	Dep	Is easily frustrated.
14	Som	Complains of pain.
15	Fun	Communicates clearly. (r)
16	Anx	Is nervous.
17	Adt	Adjusts well to changes in routine. (r)
18	Cnd	Gets into trouble.
19	Led	Gives good suggestions for solving problems. (r)
20	Hyp	Disrupts other children's activities.
21	Som	Complains about health.
22	Atn	Listens to directions. (r)
23	Dep	Is easily upset.
24	Skl	Gets along well with other. (r)
25	Anx	Worries.
26	Agg	Loses temper too easily.
27	Atn	Has trouble concentrating.
28	Adt	Recovers quickly after a setback. (r)
29	Dep	Says, "Nobody likes me."
30	Adl	Sets realistic goals. (r)

* r = reversed item

*Dep = Depression, Anx = Anxiety, Som = Somatization, Hyp = Hyperactivity, Cnd = Conduct Problems, Agg = Aggression, Atn = Attention Problems, Fun = Functional Communication, Led = Leadership, Skl = Social Skills, Adl = Activities of Daily Living, Adt = Adaptability

Table 5. Cut score selection for BASC-2 TRS screener, Child Sample

Cut Score	Sensitivity	Specificity
15	.964	.428
16	.955	.465
17	.955	.493
18	.955	.523
19	.946	.539
20	.937	.552
21	.928	.570
22	.919	.586
23	.892	.638
24	.883	.673
25	.865	.711
26	.838	.741
27	.811	.764
28	.766	.776
29	.739	.792
30	.712	.812
31	.703	.830
32	.685	.857
33	.649	.863
34	.631	.879
35	.595	.899

Table 6. BASC-2 Gate 1 TRS screener – Child Sample ROC Curve Indices

	Diagnosed	Not Diagnosed	
Positive Screen	True Positives = 90	False positives = 117	PPV = .435
Negative Screen	False Negatives = 21	True Negatives = 378	NPV = .947
	Sensitivity = .811	Specificity = .764	Overall Hit Rate = .772

Table 7. Cut score selection for BASC-2 TRS screener, Adolescent Sample

Cut Score	Sensitivity	Specificity
10	.950	.350
11	.950	.388
12	.942	.418
13	.935	.451
14	.935	.494
15	.921	.544
16	.892	.577
17	.885	.608
18	.885	.634
19	.849	.659
20	.820	.688
21	.791	.718
22	.770	.730
23	.755	.749
24	.727	.759
25	.691	.780
26	.676	.806
27	.647	.827
28	.612	.842
29	.590	.865
30	.568	.879

Table 8. BASC-2 Gate 1 TRS screener – Adolescent Sample ROC Curve Indices

	Diagnosed	Not Diagnosed	
Positive Screen	True Positives = 114	False positives = 180	PPV = .388
Negative Screen	False Negatives = 25	True Negatives = 397	NPV = .941
	Sensitivity = .820	Specificity = .688	Overall Hit Rate = .714

Table 9. Cut score selection for BASC-2 PRS screener, Child Sample

Cut Score	Sensitivity	Specificity
25	.928	.570
26	.928	.620
27	.928	.646
28	.901	.681
29	.892	.715
30	.883	.745
31	.874	.768
32	.865	.792
33	.838	.812
34	.829	.832
35	.784	.853
36	.775	.867
37	.730	.881
38	.712	.907
39	.694	.917
40	.667	.927
41	.631	.945
42	.604	.949
43	.532	.966
44	.477	.976
45	.450	.978

Table 10. BASC-2 Gate 1 PRS screener – Child Sample ROC Curve Indices

	Diagnosed	Not Diagnosed	
Positive Screen	True Positives = 93	False positives = 93	PPV = .500
Negative Screen	False Negatives = 18	True Negatives = 402	NPV = .957
	Sensitivity = .838	Specificity = .812	Overall Hit Rate = .817

Table 11. Cut score selection for BASC-2 PRS screener, Adolescent Sample

Cut Score	Sensitivity	Specificity
20	.964	.501
21	.964	.532
22	.964	.565
23	.957	.593
24	.950	.629
25	.942	.659
26	.906	.685
27	.871	.709
28	.849	.737
29	.827	.764
30	.813	.775
31	.784	.804
32	.755	.820
33	.719	.847
34	.705	.863
35	.698	.877
36	.669	.886
37	.633	.903
38	.576	.910
39	.532	.919
40	.504	.931

Table 12. BASC-2 Gate 1 PRS screener – Adolescent Sample ROC Curve Indices

	Diagnosed	Not Diagnosed	
Positive Screen	True Positives = 113	False positives = 130	PPV = .465
Negative Screen	False Negatives = 26	True Negatives = 447	NPV = .945
	Sensitivity = .813	Specificity = .775	Overall Hit Rate = .782

Table 13. PRS and TRS Gate 1 Screener Indices

Screener & Sample	Cut-off	AUC	Sensit-ivity	Specif-icity	PPV	NPV	Hit rate	Odds Ratio	Ward Statistic	Naglekerke's R^2
TRS child	27	.863	.811	.764	.435	.947	.772	13.85	98.772, p=.000	.312
PRS child	33	.891	.838	.812* Z=2.05	.500* Z=2.27	.957	.817	22.33	121.281, p=.000	.400
TRS adol	20	.829	.820	.688	.388	.941	.714	10.06	93.731, p=.000	.251
PRS adol	30	.881* Z=2.63	.813	.775* Z=3.73	.465* Z=2.96	.945	.782* Z=2.97	14.94	127.765, p=.000	.333

*statistically significant improvement from TRS (same sample) at p=.05 level

Table 14. Second Gate Full BASC-2 Cut Score Selection

Screener and Sample	Gate 2 measure	Cut off	# of False (+) w/ 2nd gate	# of False (-) w/ 2nd gate	Sens.	Spec.	PPV	NPV	Hit Rate
TRS child	BASC-2 TRS C	60/40	114	0	1.00	.026	.44	1.00	.449
		65/35	101	6	.933	.137	.454	.727	.483
		70/30	71	21	.767	.393	.493	.687	.556
TRS adol	BASC-2 TRS A	60/40	169	3	.974	.061	.396	.786	.415
		65/35	121	12	.895	.328	.457	.831	.548
		70/30	77	32	.719	.572	.512	.763	.629
PRS child	BASC-2 PRS C	60/40	92	0	1.00	.011	.503	1.00	.505
		65/35	78	0	1.00	.161	.544	1.00	.581
		70/30	49	12	.871	.473	.623	.786	.672
PRS adol	BASC-2 PRS A	60/40	127	0	1.00	.023	.471	1.00	.477
		65/35	104	7	.938	.200	.505	.788	.543
		70/30	66	23	.797	.492	.577	.736	.634
TRS child	BASC-2 PRS-C	60/40	73	5	.944	376	.538	.898	.623
		65/35	51	8	.911	.564	.617	.892	.715
		70/30	29	20	.778	.752	.707	.815	.763
TRS adol	BASC-2 PRS-A	60/40	123	2	.983	.317	.477	.966	.575
		65/35	80	11	.904	.556	.563	.901	.691
		70/30	47	25	.781	.739	.654	.842	.755
PRS child	BASC-2 TRS-C	60/40	62	5	.946	.333	.587	.861	.640
		65/35	44	15	.839	.527	.639	.766	.683
		70/30	26	32	.656	.720	.701	.677	.688
PRS adol	BASC-2 TRS-A	60/40	83	8	.929	.362	.559	.855	.626
		65/35	60	18	.841	.539	.613	.796	.679
		70/30	36	36	.681	.723	.681	.723	.704

Table 15. Effect on False Positives and Negatives of Adding a BASC-2 Comprehensive Rating Scale as Second Gate

Gate 1 Screener and Sample	Gate 2 measure	# of False (+) w/ Gate 1 only	# of False (+) w/ 2nd gate	Decrease in False (+) w/ 2nd gate	% decrease in False (+) w/ 2nd gate	# of False (-) w/ Gate 1 only	# of False (-) w/ 2nd gate	Total # of False (-) across both gates
TRS child	BASC-2 TRS C	117	71	46	39%	21	21	42
TRS adol	BASC-2 TRS A	180	77	103	57%	25	32	57
PRS child	BASC-2 PRS C	93	49	44	47%	18	12	30
PRS adol	BASC-2 PRS A	130	66	64	49%	26	23	49
TRS child	BASC-2 PRS C	117	29	88	75%	21	20	41
TRS adol	BASC-2 PRS A	180	47	133	74%	25	25	50
PRS child	BASC-2 TRS C	93	26	67	72%	18	32	50
PRS adol	BASC-2 TRS A	130	36	94	72%	26	36	62

Table 16. Logistic Regression Indices when BASC-2 Comprehensive Teacher or Parent Rating Scale is Added as Second Gate

Screener and Sample	Gate 2 measure	Odds Ratio	Ward Statistic	Naglekerke's R^2	Total R^2 across both gates
TRS child	BASC-2 TRS C	2.13	5.829, p=.016	.039	.351
TRS adol	BASC-2 TRS A	3.470	22.945, p=.000	.108	.359
PRS child	BASC-2 PRS C	6.061	23.390, p=.000	.183	.583
PRS adol	BASC-2 PRS A	3.79	20.832, p=.000	.119	.452
TRS child	BASC-2 PRS C	10.62	50.692, p=.000	.338	.650
TRS adol	BASC-2 PRS A	10.074	66.673, p=.000	.321	.572
PRS child	BASC-2 TRS C	4.912	25.078, p=.000	.181	.581
PRS adol	BASC-2 TRS-A	5.59	37.365, p=.000	.207	.540

Table 17. Comparison of Effects on False Positives and Negatives of Adding a BASC-2 Comprehensive Teacher or Parent Rating Scale as Second Gate: Same versus Different Informant

Gate 1 Screener and Sample	# of False (+) w/ same informant 2nd gate	# of False (+) w/ different informant 2nd gate	Difference in #of False (+) bet same and diff. informants	% decrease in False (+) when using diff. informant	# of False (-) w/ same informant 2nd gate	# of False (-) w/ different informant 2nd gate	Difference in # of False (-) bet same and diff. informants	% decrease in False (-) when using diff. informant
TRS child	71	29	42	59%	21	20	1	5%
TRS adol	77	47	30	39%	32	25	7	22%
PRS child	49	26	23	47%	12	32	-20	-63%
PRS adol	66	36	30	46%	23	36	-13	-36%

Table 18. Comparison of Effects on ROC Curve Indices of Adding a BASC-2 Comprehensive Teacher or Parent Rating Scale as Second Gate: Same versus Different Informant

Gate 1 Screener and Sample	Sens. (same inf)	Sens. (diff inf)	Spec. (same inf)	Spec (diff inf)	PPV (same inf)	PPV (diff inf)	NPV (same inf)	NPV (Diff inf)	HR (same inf)	HR (diff inf)
TRS child	.767	.778	.393	.752* Z=7.92	.493	.707* Z=4.55	.687	.815* Z=3.05	.556	.763* Z=4.55
TRS adol	.719	.781	.572	.739* Z=4.33	.512	.654* Z=3.53	.763	.842* Z=2.42	.629	.755* Z=3.34
PRS child	.871	.656	.473	.720* Z=5.02	.623	.701	.786* Z=2.39	.677	.672	.688
PRS adol	.797* Z=2.94	.681	.492	.723* Z=5.37	.577	.681* Z=2.39	.736	.723	.634	.704

* statistically significant at .05 level

* inf = informant, Sens. = Sensitivity, Spec. = Specificity, PPV = Positive Predictive Value, NPV = Negative Predictive Value, HR = Hit Rate

.

Table 19. Comparison of Effects on False Positives and Negatives of Adding a BASC-2 Comprehensive Teacher or Parent Rating Scale versus a BASC-2 screener as Second Gate

Gate 1 Screener and Sample	# of False (+) w/ Gate 1 only	# of False (+) w/ screener as 2nd gate	% decrease in False (+) when add Gate 2 screener	# of False (+) w/ full BASC-2 as 2nd gate	Difference in # of False (+) bet screener vs. full BASC-2	% decrease in False (+) using **full BASC-2**	# of False (-) w/ screener as 2nd gate	# of False (-) w/ full BASC-2 as 2nd gate	Difference in # of False (-) bet screener vs. full BASC-2	% decrease in False (-) using **screener**
TRS child	117	37	68%	29	8	22%	11	20	-9	45%
TRS adol	180	67	63%	47	20	30%	14	25	-11	44%
PRS child	93	37	60%	26	11	30%	14	32	-18	56%
PRS adol	130	67	49%	36	31	46%	13	36	-23	64%

Table 20. Comparison of Effects on ROC Curve Indices of Adding a BASC-2 Comprehensive Teacher or Parent Rating Scale versus a BASC-2 screener as Second Gate

Gate1 Screener and Sample	Sens. (scr)	Sens. (full)	Spec. (scr)	Spec (full)	PPV (scr)	PPV (full)	NPV (scr)	NPV (full)	HR (scr)	HR (full)
TRS child	.878* Z=2.72	.778	.684	.752	.681	.707	.879	.815	.768	.763
TRS adol	.877* Z=3.12	.781	.628	.739* Z=2.91	.599	.654	.890	.842	.725	.755
PRS child	.850* Z=3.31	.656	.602	.720* Z=2.42	.681	.701	.800* Z=2.73	.677	.726	.688
PRS adol	.885* Z=5.63	.681	.485	.723* Z=5.53	.599	.681	.829* Z=2.83	.723	.670	.704

* statistically significant at .05 level

* scr=screener, full=full BASC-2, Sens. = Sensitivity, Spec. = Specificity, PPV = Positive Predictive Value, NPV = Negative Predictive Value, HR = Hit Rate

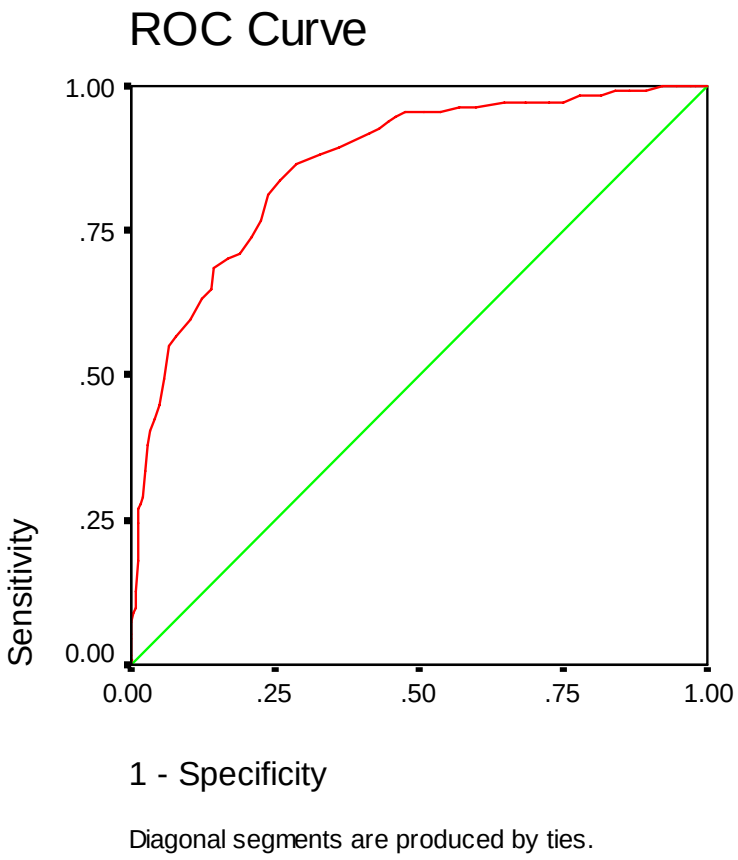

Figure 1. ROC Curve for TRS Screener, Child Sample

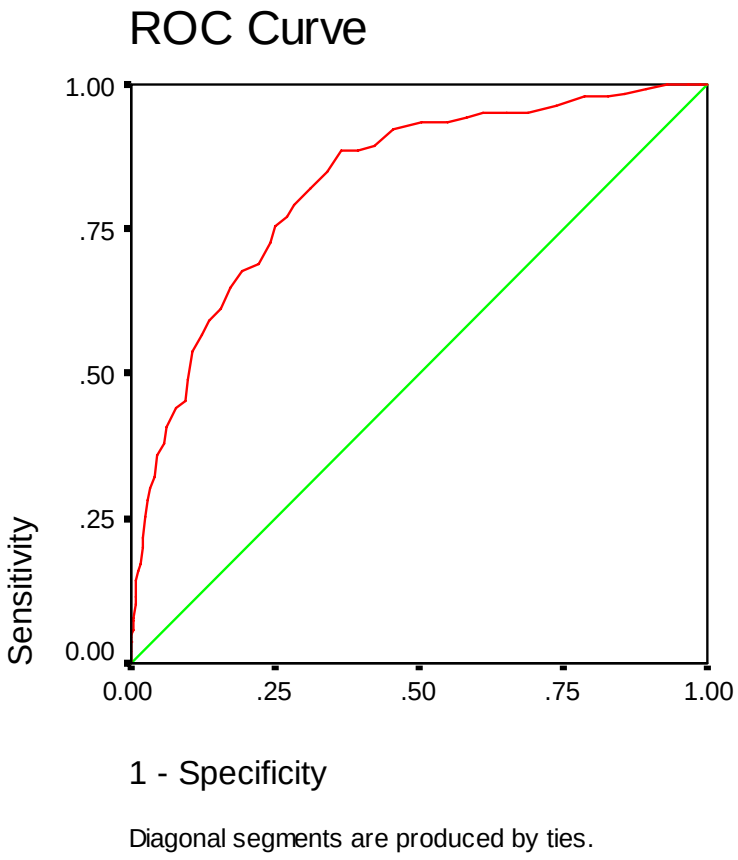

Figure 2. ROC Curve for TRS Screener, Adolescent Sample

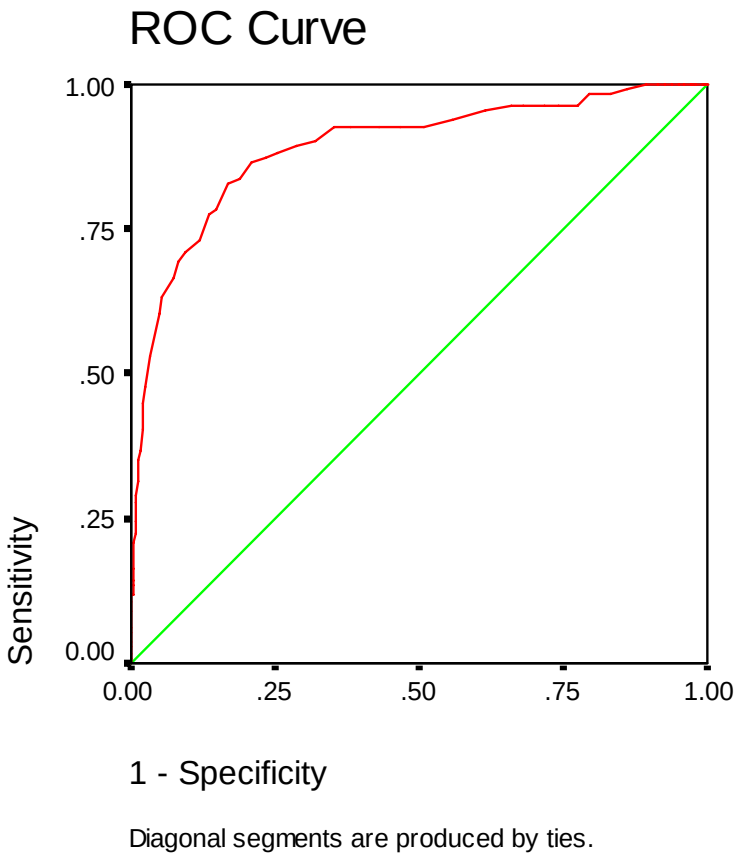

Figure 3. ROC Curve for PRS Screener, Child Sample

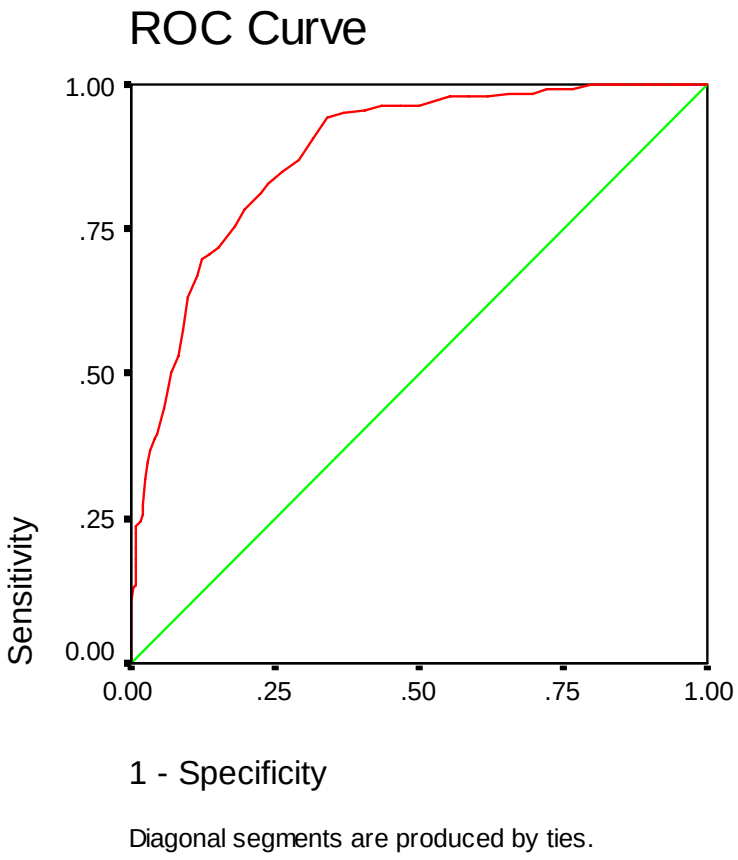

Figure 4. ROC Curve for PRS Screener, Adolescent Sample