Chapter 4
Instrumentation for Mental Health Screening

Evaluating Screening Instruments

Through universal screening, we have the potential to not only identify a greater proportion of individuals with emotional and behavioral problems, but also 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 time and money by minimizing the number of unnecessary diagnostic tests as well as reducing the length of and need for treatment and hospitalizations. 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 to determine whether these instruments have validity of score inferences, are cost-effective, and are linked with beneficial interventions and subsequent outcomes.

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." In assessing the validity of a test, the goal is not to conclude whether the test as an instrument is valid, but rather to assess the degree of validity of specific test scores for making inferences about behavior and subsequent decisions, such as intervention. Thus, validity can be viewed as an accumulation of evidence over time; it 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, known as construct validity. A construct is a latent or unobservable variable or characteristic of people that we are trying to capture as a test score or scores. 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 child's behavioral and emotional (or in medical terms "health") status and guide decision-making and intervention accordingly.

M. C. Stiffler, B. V. Dever, *Mental Health Screening at School,* Contemporary Issues in Psychological Assessment, DOI 10.1007/978-3-319-19171-3_4

Two essential steps in determining the usefulness of an instrument are assessing predictive validity, which refers to 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 (e.g., those with emotional disorders and those without such problems). By assessing these relationships, we are able to build and expand upon what is known about our proposed construct, thus continuing to accumulate evidence to support the construct validity of a measure.

When the classification of a sample of individuals is known, researchers often use an epidemiological model to determine whether an instrument can correctly classify those people as validity evidence (Derogatis and 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. In the case of mental health screening, 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 the screener identifies individuals without problems as having problems, this misclassification is referred to as the false positive rate. These types of errors may result in wasted resources and misidentification of children. False negatives occur when the screener does not identify individuals who are having problems, 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 without receiving treatment.

One can estimate the predictive power of a screener by determining the positive predictive value (PPV) and negative predictive value (NPV) (see Table 4.1). PPV indicates the proportion of individuals with positive screens who actually have emotional and behavioral problems. A low PPV indicates that 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 are present.

Table 4.1 Relationships among PPV, NPV, sensitivity, and specificity

	Diagnosed	Not diagnosed	
Positive screen	True positive (a)	False positive (b)	PPV a/a+b
Negative screen	False negative (c)	True negative (d)	NPV d/c+d
	Sensitivity a/a+c	Specificity d/b+d	Overall hit rate a+d/a+d+c+b

PPV Positive predictive value, *NPV* Negative predictive value

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 and Rosen 1955). As Hill et al. (2004) 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 0.80 and specificity of 0.95 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 et al. 2004; Campaign for Mental Health Reform 2005; Friedman et al. 1996); however, this base rate will be lower when focusing on a single disorder. Many screening instruments fail to provide PPVs and NPVs, limiting their reporting of findings to sensitivity and specificity.

Bennett and Offord (2001) have suggested that screening methods should have a minimal PPV and sensitivity of 0.50, 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 et al. (1998) considered a cut off score clinically useful if PPV or NPV was greater than or equal to 0.65 and if sensitivity or specificity was approximately 0.50 or greater. Other studies (Carran and Scott 1992; Campbell et al. 2001; Weis et al. 2005), on the other hand, indicated that sensitivity, specificity, PPV, and overall hit rate values should be equal to or greater than 0.80 to support the utility of a screening measure. For the purposes of screening, it seems that a low PPV is more tolerable than a low NPV, as false positives are more acceptable than false negatives at the stage of universal screening. Of course, the higher (closer to 1.0) all of these values are, the better the detection of the instrument. However, in the context of mental health screening, PPV values of at least 0.50 and NPV of at least 0.80 would correspond with the practical purpose of identifying as many children as possible during the screening phase, with a more focused follow-up assessment to help determine which cases were false positives.

The usefulness of a screening measure for identifying children at risk for behavioral, emotional, or academic problems can be assessed by performing a receiver operating characteristic (ROC) curve analysis to evaluate the accuracy of discrimination between children with known problems and those without. The ROC curve is a plot of the true positive rate against the false positive rate when testing different potential cut scores for a diagnostic test (Altman 1991). ROC curves demonstrate the trade-off between sensitivity and specificity: increases in sensitivity are accompanied by decreases in specificity. The area under the ROC curve is 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. An area under the curve (AUC) of 1 defines a perfect test, while an area of 0.5 represents a relatively inefficient measure; ROC curve areas of 0.80–0.90 are considered "good" discriminators while 0.90–1 are considered "excellent." Fig. 4.1 presents two ROC curves: the first for an assessment with poor discrimination (AUC = 0.64) and the second for an assessment with excellent discrimination (AUC = 0.99). The green diagonal line represents the scenario for which the

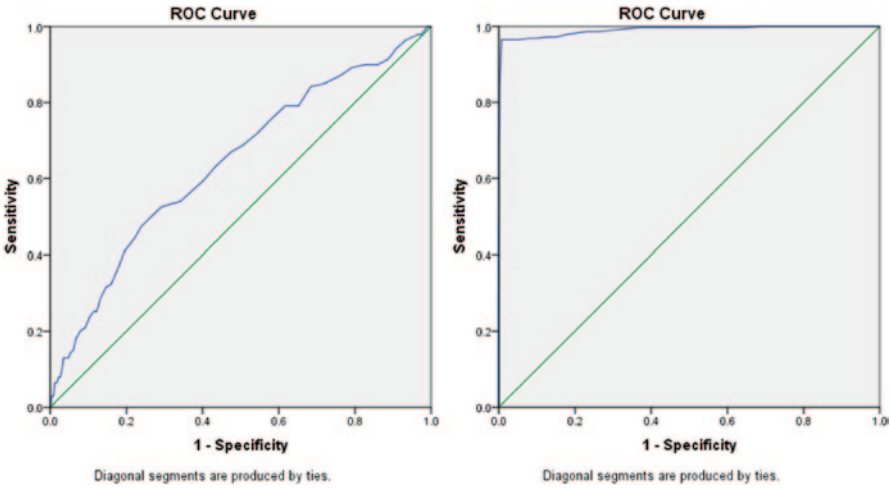

Fig. 4.1 Examples of ROC curve output for the cases of poor *(left)* and excellent *(right)* discrimination

decision to classify or not is no better than chance (0.50). The "height" of the blue curve above the diagonal is an indication of how much better an instrument is at classification as compared with flipping a coin.

Glover and Albers (2007) claimed that when evaluating screening instruments, the user should evaluate their: (a) appropriateness for intended use, (b) psychometric adequacy, and (c) usability. In this chapter, we present information on screening instruments that we hope will be a first step to selecting one to meet the goals specified by a school or district. To address appropriateness, we provide information on broadband screening measures as well as specific, single-disorder measures; the reader should select the appropriate type of instrument depending upon the constructs of interest. Externalizing problems (which may include hyperactivity/inattention), internalizing problems, and difficulties with adaptive skills represent three core constructs that are associated with mental health problems among school-age children (Frick et al. 2009); therefore, in our review we include screeners that assess difficulties in these areas broadly, with a more specific focus on disorders within the domains of externalizing and internalizing. Within each measure, we provide information related to psychometric properties and usability that we hope will assist the reader in selecting a measure with sufficient evidence for its use.

Screening Measures

The following review of child screening measures (see Table 4.2) is meant to be as comprehensive as possible; however, we do not suggest that the review is actually comprehensive as new instruments are being developed on a regular basis.

Table 4.2 Summary of available screening instruments

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Broad	Behavioral and Emotional Screening System (BASC-2 BESS)	Psychosocial risk	Parent (Pre-Kindergarten (PK)-12th) teacher (PK-12th) self-report (3rd–12th)	30 items 27 items 30 items	Nationally representative sample of 12,350 children, ages 3 through 18	Adequate reliability and validity; feasible in school settings; PPV ranges from 0.72 to 0.87 across forms and informants in prediction of BSI on BASC-2; NPV ranges from 0.94 to 0.97. Parent form: sensitivity 0.73–0.82 and specificity 0.96–0.97; teacher form: sensitivity 0.80–0.82 and specificity 0.95–0.97; self-report: sensitivity of 0.66, specificity of 0.95

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Broad	Pediatric Symptom Checklist (PSC; Jellinek et al. 1986)	Psychosocial risk	Parent (4–16) self-report (11–16)	35 items	206 children, ages 6–12 from three pediatrician's offices—99% Caucasian, Socioeconomic status (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 0.77 to 0.95 and specificity from 0.68 to 1.0; self-report: sensitivity of 0.94, specificity of 0.88 (Jellinek et al. 1995; Simonian and Tarnowski 2001; Murphy et al. 1989; Borowsky et al. 2003)
Broad	Pediatric Symptom Checklist - 17 (PSC-17; Gardner 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 0.82 and 0.81 respectively; low PPV of 0.15 (Gardner et al. 1999, 2004; Borowsky et al. 2003); more external studies needed

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
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 9577 children in the US national population sample	Adequate overall reliability and validity; reliability of specific scales is questionable; British version found sensitivity of 0.633, specificity of 0.946. PPV of 0.527, NPV of 0.964 (Goodman and Scott 1999; Mellor 2004; Goodman et al. 2003); need more research on the US version
Broad	Student Risk Screening Scale (SRSS; Drummond 1994)	Behavioral risk	Teacher (K-6; some evidence for extension to K-12)	7 items	Not available	Correlations ranging from 0.61 and 0.68 between SSRS and the SDQ total score (Lane et al. 2007). Internal consistency generally above 0.80 across grade levels; test-retest reliability generally between 0.70 and 0.86 (e.g., Lane et al. 2007, 2008, 2009, 2013; Oakes et al. 2010)

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Broad	Systematic Screening for Behavior Disorders (SSBD; Walker and Severson 1992)	Psychosocial risk	Teacher (originally K-6, extended to K-8)	Three stages: 1. Teacher ranking of all students in the classroom, 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	454 students and 18 teachers from grades 1–5 in Springfield, or (Walker et al. 1989). Nationally standardized in follow-up work, normative data by age and gender available (Walker and Severson 1992)	Evidence of discriminant validity (externalizing vs. internalizing) and concurrent validity with scales of the CBCL (Walker et al. 1989). Test-retest reliability: 0.83–0.88, internal consistency: 0.82–0.88

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Broad	Social, Academic, and Emotional Behavioral Risk Screener (SAEBRS; Kilgus et al. 2013)	Psychosocial risk	Teacher (K-12)	19 items Social: 6 items academic: 6 items emotional: 7 items	Original SABRS was developed based on data from 54 teachers who rated 243 students (K-5) in 3 elementary schools in the Southeastern US about 51 % of the students were white, 33 % were African American, and 10 % were Hispanic/Latino. 20 % of the students received special education services	Adequate reliability and convergent validity; internal consistency of 0.90–0.94. Sensitivity of 0.85–0.97, specificity of 0.73–0.84 (Kilgus et al. 2013)
Specific	Beck Youth Inventories of Emotional and Social Impairment (BYI-II; Beck et al. 2001)	Depression, anxiety, anger, disruptive behavior, and self-concept	Self-report (7–18)	Five 20-item screens	800 children (7–14) stratified based on 1999 census data on, sex, SES and ethnicity; no indication of stratification by geography	Adequate reliability and convergent validity; however discriminant validity evidence is questionable—majority of scales seem to measure same construct (Bose-Deakins and Floyd 2004)

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
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: 1286 subjects, ages 9–17, from 4 sites (Atlanta, New Haven, New York, and Puerto Rico)	Adequate reliability and validity; sensitivity of 0.68, specificity of 0.91, PPV of 0.34, NPV of 0.98 (Lucas et al. 2001; Leung et al. 2005); need more external validity studies
Specific	Conners 3rd edition (Conners 1973, 2008; Conners et al. 1998)	Oppositional, cognitive problems, hyperactivity	Parent (6–18) and teacher (6–18) and self-report (8–18)	Long forms: 100–115 items; short forms: 41–45 items; ADHD Index (10 items) DSM-IV symptom checklist (18 items)	A representative sample of children ages 6–18, based on 2000 US census data	Adequate reliability and validity; sensitivity from 0.78 to 0.92, specificity from 0.84 to 0.94, PPV from 0.83 to 0.94, and NPV from.81 to 0.92 (Connors et al. 1998); feasible in school settings; limited psychometric evidence for newest edition currently

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Specific	AD/HD Comprehensive Teacher's rating scale – 2nd edition (ACTeRS-2; Ullman et al. 1988, 1997)	ADHD	Teacher (K-8th grade)	24 items	2362 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
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 2000 children ages 4–20 matched to 1990 US. census data	Excellent reliability and validity; however, parent form has low specificity evidence; parent form: sensitivity from 0.83 to 0.84, specificity low at 0.49, PPV from 0.54 to 0.58, and NPV from 0.77 to 0.81; teacher form: sensitivity from 0.63 to 0.72. specificity at 0.86, PPV from 0.78–0.79, and NPV from 0.73 to 0.81 (DuPaul et al. 1998)

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Specific	Eyberg Child Behavior Inventory; (Eyberg and 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 0.63 to 0.96, specificity from 0.87 to >0.90, PPV from 0.63 to 0.88 and NPV from 0.82 to 0.96 (Eyberg and Robinson 1983; Boggs et al. 1990; Rich and Eyberg 2001; Weise et al. 2005)
Specific	Sutter-Eyberg Student Behavior Inventory—revised (Eyberg and 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

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Specific	Revised Children's Manifest Anxiety Scale, second edition (RCMAS−2; Reynolds and Richmond 2008)	Anxiety	Self-report (6–19)	49 items short form—10 items	2368 individuals aged 6–19, representative of the US population in terms of gender, ethnicity, and SES; stratified into three age groups: 6–8, 9–15, 15–19	Reliability estimates improved over the original version: 0.92 for total score and ranging from 0.75–0.86 for scale scores. Similar psychometric properties found for children in Singapore (Ang et al. 2011) and Pakistan (Ahmad and Mansoor 2011)
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 et al. 1994; Southam-Gerow et al. 2003); unable to differentiate between disorders

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Specific	Multidimensional anxiety scale for children, second edition (March 2013)	Anxiety	Self-report (8–19) Parent (8–19)	50 items 50 items	Self-report normative sample includes 1800 children aged 8–19 years. The MASC 2–parent normative sample includes 1600 with an equal number of boys and girls being rated within the 8–19 age range by parents. All normative data is representative of the US and Canadian population in terms of ethnicity/race, gender, and age	Internal consistency 0.92 for self-report, 0.89 for parent-report. Test-retest reliabilities from 0.80 to 0.94 over 1–4 weeks. Studies on original MASC yielded adequate composite reliability—some subscales lower, and initial validity (March et al. 1997); more studies needed to assess discriminant validity; short form has low reliability and lacks validity evidence; limited information on new MASC-2 release to-date beyond information in the manual

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Specific	Reynolds Child Depression Scale, second edition (RCDS-2; Reynolds 2010) Reynolds Adolescent Depression Scale, second edition (RADS-2; Reynolds 2002)	Depression	RCDS-2: Self-report (7–13) RADS-2: Self-report (11–20)	30 items short form—11 items	RCDS-2: A new standardization sample of students, drawn from 11 states, stratified closely to match the US census data for gender and ethnic background. The new sample also included children in Grade 2. RADS-2: Sample of 3300 adolescents, stratified to match 2000 US census data.	RCDS-2: Internal consistency coefficients range from 0.87 to 0.91. RADS-2: Internal consistency coefficients range from 0.80 to 0.94. Four factors identified as dysphoric mood, anhedonia-negative affect, negative self-evaluation, and somatic complaints. These factors have been upheld in independent analyses (Osman et al. 2010)
Specific	Children's Depression Inventory 2 (CDI 2; Kovacs 2010)	Depression	Parent (7–17) teacher (7–17) self-report (7–17)	Parent: 17 items teacher:12 items self-report: 28 items self-report short form: 12 items	1100 children aged 7–17 years from 26 different states in the US; stratified by ethnicity, age, gender; 600 teachers, 800 parents; also includes clinical sample	Acceptable reliability and classification accuracy with total score values of: Sensitivity at 0.83, specificity of 0.73, PPV 0.76, and NPV of 0.81

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Specific	Center for Epidemiological Studies Depression Scale Modified for Children (CES-DC; Faulstich et al. 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 et al. 1986). In a recent meta-analysis, average internal consistency was 0.88, average sensitivity was 0.76, and average specificity was 0.71, average PPV values ranged from 0.08 to 0.32, and NPV values ranged from 0.12 to 0.98
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)

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Specific	Beck scale for suicidal ideation (Beck et al. 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 et al. 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 1290 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 0.83 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 0.87 to 1.0, but lower specificity ranging from 0.54 to.64 (Thompson and Eggert 1999)

Table 4.2 (continued)

Instrument type	Instrument	Conditions addressed	Informants and age/grade ranges	# of items	Norming/development sample	[a]Reliability/validity
Specific	Columbia Health/Suicide Screen (CSS; Shaffer et al. 2004)	Depression, suicide	Self-report (11–18)	14 items	Convenience sample of 1729 9th − 12th graders from 7 NY high schools	PPV is low at 12–16 %; NPV 99 %; sensitivity of 0.75 to 0.88 and specificity of 0.83 (Shaffer et al. 2004); more research needed

BESS Behavioral and Emotional Screening System, *PSC* Pediatric Symptom Checklist, *PSC—17* Pediatric Symptom Checklist—17, *SDQ* Strengths and Difficulties Questionnaire, *SRSS* Student Risk Screening Scale, *SSBD* Systematic Screening For Behavior Disorders, *SAEBRS* Social, academic, and emotional behavioral risk screener, *BDI* Beck Youth Inventories of emotional and social impairment, *Conners-3* Conners 3rd edition *DPS—4.32* DISC (Diagnostic Interview Schedule for Children) Predictive Scales, , *ACTeRS-2* AD/HD Comprehensive Teacher's Rating Scale—2nd edition, *ADHD-IV* ADHD Rating Scale-IV, *RCMAS-2* Revised Children's Manifest Anxiety Scale Second Edition, *STAIC* State-Trait Anxiety Inventory for Children, *RCDS-2* Reynolds Child Depression Scale Second Edition, *CDI* 2 Children's Depression Inventory 2**,** *CES-DC* Center for Epidemiological Studies Depression Scale Modified for Children, *CDS* Columbia Depression Scale, *SIQ-JR* Suicidal Ideation Questionnaire Jr., *SIQ* Suicidal Ideation Questionnaire, *CSS* Columbia Health/Suicide Screen

[a] For screening purposes only—not diagnostic, but rather an indication for further assessment

We focused on those measures specifically developed for elementary school-aged children that had been the subject of research studies and contain information on psychometric properties in their manuals.

Broadband Screening Measures

The existence of brief, broadband screeners may provide an important piece of the infrastructure needed to convince school districts and health care providers that early identification is not only beneficial to children, but also can be 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. A broadband screening measure would ameliorate this problem by covering a number of problem areas in one brief measure.

Theoretically, a broadband screener is feasible if one invokes modern temperament and neurological theory and their variants (Gray 1987; Rothbart and Bates 1998). Although beyond the scope of this text, 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 comorbidity is highly prevalent in child psychopathology (Rutter and Sroufe 2000). Additional support can be found in the numerous factor analytic studies of child behavior rating scales that produce three or four factor solutions (Reynolds and 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 et al. (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 (discussed in Chap. 6), screeners should simply identify those children with elevated symptomatology, leaving diagnosis of specific disorders to the later gates. Broadband 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 and Edelbrock 1987), Behavior Assessment System for Children–2 (BASC-2; Reynolds and Kamphaus 2004), McDermott Adjustment Scales for Children and Adolescents (ASCA; McDermott et al. 1994), Child/Adolescent Psychiatry Screen (CAPS), Swanson, Nolan, and Pelham Rating Scale -Revised (SNAP-IV-R; Swanson and Carlson 1994), and the McCarney Behavior Evaluation Scale—2 (McCarney and Leigh 1990).

Therefore, a need exists for the development of brief, multidisorder child screening measures of emotional and behavioral adjustment. Several of the screening instruments listed below are broad, multidisorder instruments that have potential to serve this need; however, these instruments are nascent, and more information is needed about their psychometric properties across populations and time.

Pediatric Symptom Checklist (PSC)

One measure that may be considered a true, multidisorder screener is the PSC (Jellinek et al. 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 4–16). This measure has been extensively studied with a range of economically, racially, and clinically diverse samples and has been 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 et al. 1986, 1995; Jellinek and Murphy 1988; Murphy et al. 1992; Simonian and Tarnowski 2001; Stoppelbein et al. 2005; Walker et al. 1989). It has been found to have good sensitivity, ranging from 0.77 to 0.95, and specificity, ranging from 0.68 to 1.0 (Jellinek et al. 1995; Jellinek and Murphy 1990; Simonian and Tarnowski 2001; Stoppelbein et al. 2005; Walker et al. 1989). 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. The PSC has also identified students whose difficulties were unknown to school staff, thus suggesting that it may be of use in school settings as well (Murphy et al. 1989). However, a teacher version of this instrument does not currently exist.

Pagano et al. (2000) adapted the PSC into self-report format (PSC-Y) 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. 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 et al. (2000) found support for the use of the PSC-Y in a high school-based health center environment as well. It demonstrated acceptable levels of sensitivity (0.94) and specificity (0.88) in identifying children at psychosocial risk (Pagano et al. 2000); PPV and NPV were not reported. However, the AUC of the PSC-Y was 0.66, which is lower than the 0.8 needed to be considered satisfactory. Therefore, caution should be used in making classification decisions based on the PSC-Y.

Gardner et al. (1999) created a short form of this instrument, Pediatric Symptom Checklist—17 (PSC-17), which has demonstrated lower preliminary reliability estimates at 0.67 for the total score (Borowsky et al. 2003). This instrument has been found to have adequate sensitivity at 0.82 and specificity at 0.81; however, its PPV was found to be quite low at 0.15 (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 Behavioral and Emotional Screening System (BESS)

In 2007, Kamphaus and Reynolds developed the BASC-2 Behavioral and Emotional Screening System (BESS), a multi-informant screening system focused on detecting risk for the development of a disorder, rather than any specific diagnosis. The BESS was developed such that item content would reflect the major constructs of child adjustment as contained within the full BASC-2 rating scales. Factor analyses suggest that the self-report includes the domains of internalizing problems, inattention, school problems, and adaptive skills (Dowdy et al. 2011a), the teacher report includes internalizing problems, externalizing problems, school problems, and adaptive skills (Dever et al. 2012), and the parent report includes internalizing problems, externalizing problems, inattention, and adaptive skills (Dowdy et al. 2011b).

The BASC-2 BESS includes two teacher forms (Preschool for ages 3 through 5, and Child/Adolescent for Grades K through 12), two parent forms (Preschool for ages 3 through 5 and Child/Adolescent for Grades K through 12), and a student self-report form (Grades 3 through 12). All forms contain between 25 and 30 items and take 5–10 minutes to administer. All items are rated on a 4-point scale (i.e., *never, sometimes, often, almost always*). A raw score is created by summing the responses to the problem items and the reverse scores of the adaptive behavior items. The raw score is transformed to a total *T*-score, in which higher scores reflect more problems; 20–60 suggests a "Normal" level of risk, 61–70 suggests "Elevated" risk, and scores of 71 or higher suggest an "Extremely Elevated" level of risk.

Reliability evidence was excellent; all split-half reliability coefficients were greater than 0.90 and test-retest reliabilities ranged from 0.80 to 0.91 (Kamphaus and Reynolds 2007). Preliminary validity evidence is also strong; the manual presents strong correlations with other emotional and behavioral measures, the ability to predict important school outcomes, including academic performance, and adequate ROC curve indices (Kamphaus and Reynolds 2007). Interrater reliability estimates ranged from 0.71 to 0.80 for teachers, and from 0.82 to 0.83 for parents. Dever et al. (2013) provided evidence that the BESS screener can provide useful mental health surveillance information across schools and districts, in addition to the individual data gathered. More research, especially validity studies focusing on different outcomes and diverse samples, must be conducted on these new instruments to adequately assess their validity.

Strengths and Difficulties Questionnaire (SDQ)

The SDQ is a 5 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 4–16-year olds and also includes a self-report version for 11–16-year olds. The SDQ was developed in Great Britain based on theory using Diagnostic and Statistical Manual of Mental Disorders-IV (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 (Becker et al. 2004; Flawes and Dadds 2004; Goodman 2001; Malmberg et al. 2003; Ronning et al. 2004; Van Widenfelt et al. 2003; Vostanis 2006).

In several countries, the total score has been found to have adequate reliability with an internal consistency of 0.76 and test-retest reliability of 0.96; however, the internal consistency of the individual scales, with the exception of the inattention-hyperactivity scale, has been questionable. This is especially true for Peer Problems which has an alpha of 0.51 (Goodman and Scott 1999; Mellor 2004). In 2003, Goodman, Ford, Simmons, Gatward, and Meltzer performed a ROC curve analysis on a British community sample of 7984 5–15 year olds using the SDQ and found a sensitivity of 0.633, specificity of 0.946, PPV of 0.527, and NPV of 0.964. Sensitivity varied by diagnosis with 70 to 90% of conduct, hyperactivity, depression, developmental disorders, and some anxiety disorders being identified, but only 30–50% of those children with specific phobias, panic disorders, eating disorders, and separation anxiety being identified. They found that the 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.

Lane et al. (2012a) provide a chapter summarizing the psychometric properties and use of the SDQ. They highlight the low false positive rate found in studies by Goodman and colleagues, but the higher false negative rate may be concerning in a universal screening program due to the desire to identify as many children who may need supports as possible. In this chapter, the authors provide several examples of schools that have used the SDQ to inform intervention and prevention efforts from preschool through high school. Additional information regarding the feasibility of use of the SDQ at the preschool level is presented in White et al. (2013).

Taken together, this research suggests that the SDQ (and more specifically, the total score) would be best used as an indicator of general maladjustment with a second step 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 it is critical to identify 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 (as discussed further in Chap. 6).

An American version of the SDQ has been developed more recently and preliminary findings are positive (Bourdon et al. 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. 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.

Systematic Screening for Behavior Disorders

Systematic Screening for Behavior Disorders (SSBD; Walker and Severson 1992) is a multiple-gated procedure developed to identify students in elementary school who are at elevated risk for externalizing or internalizing disorders. Since its initial development, the SSBD has been extended to the middle school grades as well (e.g., Caldarella et al. 2008). This screening procedure consists of three stages. At stage 1, teachers are asked to create two "top 10 lists"—one for students with internalizing issues and another for students with externalizing issues. In this procedure, teachers are instructed that no student can appear on both lists. The top three students on each list (6 in total) continue to stage 2. During stage 2, teachers complete two rating scales for each of these six students, which capture both the behaviors of those children and the frequency or intensity of those behaviors. Any students exceeding the normative criteria on these instruments continue to stage 3. At stage 3, the students who were identified at the end of stage 2 are observed by a trained professional (often a behavioral specialist or school psychologist) both in the classroom and on the playground. Data on engagement and social behavior are recorded in order to frame the results of the rating scales and assist with decisions about intervention or referral.

Initial development of the SSBD yielded stability coefficients ranging from 0.83 to 0.88, and internal consistency coefficients ranging from 0.82 to 0.88 (Walker et al. 1988). In addition, Walker et al. (1990) provided criterion validity evidence for the SSBD based on school records of behavior and special education classifications. Finally, there is sufficient evidence of convergent validity with similar measures, including the SSRS (Lane et al. 2009) and the CBCL (Walker et al. 1988).

Student Risk Screening Scale

The Student Risk Screening Scale (SRSS; Drummond 1994) is a 7-item teacher-report instrument designed to detect risk for behavior problems in grades K-6; more recent research has provided evidence that its use can be extended to grades K-12

(Lane et al. 2008). All items are rated on a scale from 0 to 3, for a total possible score of 21. Based on their total scores, students are categorized into three levels of risk: Low (0–3), Moderate (4–8), and High (9–21). Due to its brevity, teachers can complete the assessment for an entire classroom in approximately 15 minutes, making it a practical choice for universal screening via teacher-report. The SRSS is available both in written and electronic forms.

In terms of validity evidence, Lane et al. (2009) compared the SRSS and the SSBD (Walker and Severson 1992) at the Kindergarten through third grade levels. Students in this study were enrolled in seven elementary schools, and were predominantly White (95%). SRSS scores were used to predict SSBD risk classification. The SRSS performed similarly to the SSBD at identifying externalizing problems, but performed poorly at identifying children with internalizing problems. However, this aligns well with the original purpose of the SRSS to assess for problems related to antisocial behavior. In addition, among a diverse group of elementary school students, Menzies and Lane (2012) found that SRSS scores predicted the number of office disciplinary referrals a child would receive during an academic year.

Among older groups of students, Lane et al. (2007) found correlations ranging from 0.61 and 0.68 between SRSS and the SDQ total score for middle school students. There is evidence of adequate internal consistency, test–retest stability, and predictive validity (using the criteria of grade point averages, office disciplinary referrals, and out of school suspensions) of the SRSS for use among urban middle school students (Lane et al. 2010). There is also evidence supporting the use of the SRSS among high school students (Lane et al. 2008). Internal consistency and test-retest reliability coefficients are similarly high across grade levels (see Table 4.2).

More recent efforts have adapted the SRSS to include items that assess internalizing difficulties as well, yielding the Student Risk Screening Scale for Internalizing and Externalizing Behaviors (SRSS-IE; Lane et al. 2012b). Although the original SSRS-IE included the original 7 items of the SRSS plus an additional 7 internalizing items, initial factor analytic work among over 2000 students in grades K-6 supported the retention of only 5 internalizing items, for an SRSS-IE scale of 12 items in total (Lane et al. 2012). Preliminary convergent validity evidence suggests that the SRSS-IE predicts both SDQ and SSBD scores among this elementary school sample. Furthermore, the development of an 11-item Student Risk Screening Scale for Early Childhood (SRSS-EC; Lane et al. 2015) has shown initial promise for identifying the internalizing and externalizing difficulties of preschool students.

Social, Academic, and Emotional Behavioral Risk Screener

The social, academic, and emotional behavioral risk screener (SAEBRS; Kilgus et al. 2013) is a 19-item teacher-report screening instrument that consists of three domains: social behavior (6 items), academic behavior (6 items), and emotional behavior (7 items). Students are rated on a 4-point Likert-type scale, from 0 (never) to 3 (almost always). The SAEBRS can be completed in less than 3 minutes per student, and is intended for rating students in grades K-12. Users of the SAEBRS are provided with an overall level of risk for each student rated, as well as risk levels within each of the three domains of interest.

Factor analysis work with the original dual factor SABRS 12-item instrument (prior to the addition of the 7 emotional behavior items) supports the structure of one broad factor (Behavior) under which are two narrow factors (social and academic) at both the elementary (Kilgus et al. 2013) and secondary (Kilgus et al. 2015) grade levels. Across grade levels, internal consistency estimates are similarly high (ranging from 0.89 to 0.94). When multiple teachers rate the same high school student, interrater reliability estimates range from 0.35 to 0.51 (Kilgus et al. 2015). Future research is needed to determine how the addition of the emotional behavior items has changed the psychometric properties of the instrument. Also, longitudinal research is necessary to examine the predictive validity of the SAEBRS in regard to important social, emotional, and academic outcomes of interest.

Specific Screeners for Multiple Disorders

Several child emotional and behavioral screeners consist of a number of quick screens for multiple disorders simultaneously. For example, the *Beck Youth Inventories—Second Edition* (BYI-II; Beck et al. 2005) are designed for children ages 7–18 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.

The DISC (Diagnostic Interview Schedule for Children) Predictive Scales—version 4.32 (DPS-4.32; Leung et al. 2005) was updated to include work done on the National Institute of Mental Health (NIMH) DISC-IV (Shaffer et al. 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 8–18, having any of 18 disorders. Additionally, the DPS-4.32 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 et al. 1996), by identifying those items that were most predictive of specific diagnoses (Lucas et al. 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 et al. (2001) examined the DPS-2.3 classification accuracy for a number of disorders including simple phobia, social phobia, agoraphobia, Obsessive Compulsive Disorder (OCD), Major Depressive Disorder (MDD), attention deficit hyperactivity disorder (ADHD), ODD, and conduct disorder. They found adequate reliabilities, sensitivities ranging from 0.67 to 1.00, specificities from 0.49 to 0.96, PPV from 0.07 to 0.74, and NPV from 0.87 to 1.00. They concluded that the DPS is a valuable tool for determining subjects who do not need further assessment and for 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 reli-

ability as well as adequate specificity (0.91), and NPVs (0.98); however, sensitivity was a bit low at 0.68 and PPV was found to be 0.34. Once again, more research should be done to reinforce these findings on other samples (Leung et al. 2005).

Specific Screeners

Other child emotional and behavioral screeners tend to focus on one or several specific diagnoses or problems. These screeners can be classified as those with a focus on specific externalizing disorders, and those with a focus on specific internalizing disorders or risk for suicide. Below we review some of the available screening instruments in each category.

Externalizing Measures

Externalizing disorders, especially ADHD, have been the focus of numerous screening measures for children. The Conners 3rd Edition (CRS-3; Conners 1973, 2008; Conners et al. 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 Conners-3 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 (ages 6 through 18), teacher (ages 6 through 18), and adolescent (ages 8 through 18) self-report—all of which also have short (10 minutes) and long (20 minutes) forms available. The long forms are too extensive to be used as screening measures; however, in addition to short forms of the Conners-3, users also have the option of administering a 10-item ADHD index or the brief DSM-IV and Diagnostic and Statistical Manual for Mental Disorders-V (DSM-V) Symptom Scales.

Previous versions of this instrument have been found to have adequate reliability and validity (e.g., Conners' Rating Scales–Revised (CRS-R); Conners et al. 1997), but were criticized for having too low cutoff scores thus inflating prevalence rates. However, classification indices are quite high with sensitivities ranging from 0.78 to 0.92, specificities ranging from 0.84 to 0.94, PPV ranging from 0.83 to 0.94, and NPV ranging from 0.81 to 0.92 depending on informant (parent, teacher, and adolescent) (Conners et al. 1997). To date, the Conners-3 manual is the best source of psychometric information for these scales. Alpha coefficients ranged from 0.84 to 0.97 across subscales and informants, indicated adequate internal consistency. Test-retest reliabilities ranged from 0.71 to 0.98, and values for interrater reliabilities ranged from 0.74 to 0.94 for the parent form and from 0.52 to 0.82 for the teacher form.

The ADHD Comprehensive Teacher's Rating Scale (ACTeRS-2**;** Ullman et al. 1988, 1997) is a 24-item teacher-rated ADHD screener created using a normative

sample of over 3700 children from kindergarten through 8th grade. The instrument produces four subscales: attention, hyperactivity, social skills, and opposition. 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 specific information regarding the standardization sample. Ullman et al. (2000) found that the ACTeRS could differentiate between children with and without ADHD as well as children with learning disabilities and those with ADHD. Although it has not been validated as a screening measure, the ACTeRS-2 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.

In a study including students from grades K through 5 in the mid-Atlantic US, Erford and Hase (2006) found adequate internal consistency of the ACTeRS-2 subscales (from 0.89 to 0.93); however, factor analyses in this same study supported a two-factor solution rather than the four factors suggested by the authors of the instrument. The 30-day test-retest reliabilities ranged from 0.80 to 0.89. When compared to a diagnosis from a qualified mental health professional, 83 % of those diagnosed as ADHD-inattentive type (sensitivity: 0.77; specificity: 0.88), and 0.86 % of those diagnosed as ADHD-hyperactive/impulsive type (sensitivity: 0.81; specificity: 0.88) were correctly identified.

The ADHD Rating Scale—IV (ADHD-IV; DuPaul et al. 1998) is an 18-item rating scale for children ages 5–18, 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, divergent, and predictive) evidence (DuPaul et al. 1998). The manual also provides different cutoff scores depending on the purpose of the assessment (rule-out/screening vs. diagnosis). Parent ratings have sensitivities of 0.83–0.84, specificities of 0.49, PPV of 0.54–0.58, and NPV of 0.77–0.81. Teacher ratings produce sensitivities of 0.63–0.72, specificities of 0.86, PPV of 0.78–079, and NPV of 0.73–0.81 (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 et al. (2003) warn users about the risk of misclassifying youth due to suboptimal sensitivity and specificity. Furthermore, the scale has yet to be updated to match DSM-V diagnostic criteria, and the lack of a self-report may be a limitation for certain contexts and applications.

The Eyberg Child Behavior Inventory (ECBI; Eyberg and 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 and 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 and served to replace non-school related items from the ECBI. The standardization of the SESBI-R is problematic; the norming sample consisted of 415 elementary school children from 11 schools in Gainesville, FL but the SESBI-R is supposed to target children ages 2–16 despite being normed on a more narrow aged group of children (Meikamp 2003).

The ECBI has been found to have adequate reliability and concurrent validity (Boggs et al. 1990). The ECBI was also found to discriminate between normal and conduct-problem adolescents (Eyberg and Robinson 1983). Rich and Eyberg (2001) found the ECBI to have excellent classification accuracy in a sample of children ages 3–6 with a sensitivity of 0.96, specificity of 0.87, PPV of 0.88, indicating that 88 % of the children who exceeded the cutoff score were correctly identified, and NPV of 0.96.

Weis et al. (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 0.63 for the Conduct problem component of the ECBI to 0.77 for the Inattentive component. Specificities were all above 0.90. They found that all components of the ECBI displayed adequate NPV, ranging from 0.82 to 0.94. The ECBI Inattentive and Oppositional components displayed PPV of 0.85 and 0.80 respectively, while the conduct problem component exhibited lower PPV at 0.63. The SESBI-R has some preliminary reliability and validity evidence; however, no reliability or validity evidence exists for older children (Whiston and Bouwkamp 2003). More research is needed on the SESBI-R.

Internalizing Measures

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, Second Edition (RCMAS-2; Reynolds and Richmond 2008), the State-Trait Anxiety Inventory for Children (STAIC; Spielberger 1973), the Multidimensional Anxiety Scale for Children, Second Edition (MASC-2; March 2013), the Reynolds Child Depression Scale, Second Edition (RCDS-2; Reynolds 2010), and the Children's Depression Inventory-2 (CDI-2; Kovacs 2010).

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 those without any disorders; however, findings are mixed on their ability to discriminate among diagnostic groups, especially between internalizing problems such as anxiety and depression (Kamphaus and Frick 2002; Seligman et al. 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.

In 2008, an updated second edition of the RCMAS was developed. This edition has an updated standardization sample, improved psychometric properties with improved reliability over the original version, and additional items meant to expand content coverage and reflect changes in the way children now experience anxiety

(Reynolds and Richmond 2008). Internal consistency of the total score was reported as 0.92 with a test–retest reliability of 0.76. The scale consists of four factors: physiological anxiety, worry, social anxiety, and defensiveness; however, the total score has yielded higher reliability estimates than the factor scores, which should be used with caution in the absence of more psychometric testing (Huberty 2012). Additionally, a short form consisting of the first ten items of the full form was added that yields a short form total anxiety score. The manual suggests that this form would be useful when screening large numbers of children.

The MASC-2 (March 2013) is a recently updated multi-rater anxiety measure with 50-item self-report (MASC 2–SR) and parent (MASC 2–P) rating forms developed for youth aged 8–19. The MASC-2 yields a total score and subscale scores for six disorder-specific areas. Internal consistency estimates are good, at 0.92 for the MASC 2-SR total score and 0.89 for the MASC 2-P total score. However, internal consistency estimates are lower for the individual subscales (median 0.79). Test–retest reliabilities for both forms range from 0.80 to 0.94. Inter-rater reliability across the two informant forms ranged from 0.43 to 0.68. It has been suggested that the MASC 2-SR might be especially useful for school-based screening, as it can be administered by teachers in an RtI model (Fraccaro et al. 2015). The original version of the MASC (March et al. 1997) has been found to have adequate reliability, including test–retest reliability (March and Sullivan 1999; Christopher 2001), as well as good convergent and divergent validities (March et al. 1997). Rynn et al. (2006) used the MASC to discriminate between children with generalized anxiety disorder and children with depression. They found the AUC of 0.623 to be in the poor to fair range. When sensitivity was set at 0.80, maximum specificity was found to be 0.34. This instrument has not been validated as a screening instrument in a multiple-gate screening system. To date, there is limited information on the revised MASC-2.

The CDI-2 (Kovacs 2010) is a revision of the original CDI(Kovacs 1992) that includes new items that focus on the core aspects of childhood depression, revised scales, and new norms that are representative of the US population. The CDI-2 is a comprehensive multi-rater assessment of depressive symptoms in youth aged 7–17 years. It consists of a 28-item self-report form that yields a total score, two scale scores (emotional problems and functional problems), and four subscale scores, a short self-report form that contains 12 items and yields a total score, as well as teacher and parent forms. Self-report items are answered on a 3-point scale, whereas parent- and teacher-report items are answered on a 4-point scale. The correlation between the CDI-2 self-report and self-report short-form was found to be 0.95, indicating that the short-form may prove quite useful for efficient screening. The manual reported acceptable reliability and classification accuracy with total score values of: sensitivity of 0.83, specificity of 0.73, PPV 0.76, and NPV of 0.81.

The RCDS-2 (Reynolds 2010) is another self-report measure intended to assess the severity of depressive symptomatology in children ages 7–13. The RCDS-2 retains the 30 items used in the original measure, but presents updated normative data. It also includes a short form consisting of 11 of the most critical items from full form. Internal consistency coefficients are satisfactory, ranging from 0.87 to 0.91. For the original RCDS, sensitivity of 0.73 and specificity of 0.97 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 and Frick 2002). It is advertised for use as a large-scale screening instrument. The Reynolds Adolescent Depression Scale-Second Edition (RADS-2; Reynolds 2002) was designed as a self-report form for informants aged 11–20. Internal consistency coefficients range from 0.80 to 0.94. It contains four factors identified as dysphoric mood, anhedonia-negative affect, negative self-evaluation, and somatic complaints. These factors have been upheld in independent analyses (Osman et al. 2010).

The Center for Epidemiological Studies Depression Scale Modified for Children (CES-DC; Faulstich et al. 1986) was adapted from the adult CES-D. Faulstich and colleagues (1986) found that the measure had poor reliability and validity for children. A recent meta-analysis, on the other hand, suggests that the measure has adequate reliability and validity for children and adolescents (Stockings et al. 2015). Across nine studies, the average internal consistency was found to be 0.88, average sensitivity was 0.76, and average specificity was 0.71. However, average PPV values ranged from 0.08 to 0.32, and NPV values ranged from 0.12 to 0.98, indicating the need for further evidence of the use of this tool as a screening instrument among children and adolescents. 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 DISC (Shaffer et al. 2000); however, this scale is lacking reliability and validity evidence (Table 4.2).

Suicide Measures

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 15–24. 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 et al. 2004). As Shaffer et al. (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 et al. 1979), the Suicide Risk Screen (Eggert et al. 1994), and the Suicidal Ideation Questionnaire (Reynolds 1989), which yielded adequate sensitivity ranging from 0.83 to 1.00 with less than adequate specificity from 0.40 to 0.70 when used 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 factors found to predict suicide among adolescents 14 years and older: suicidal ideation, suicide attempts, depression, and substance use (Shaffer et al. 2004; Brent et al. 1999). Thompson

and Eggert (1999) found the Suicide Risk Screen to have sensitivity ranging from 0.87 to 1.00, but low specificity from 0.54 to 0.64 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 among youth ages 11–18. 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 et al. (2004) found this instrument to have adequate sensitivity (0.75) in identifying high school students 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 0.83 is superior to most other instruments (p. 71). The PPV was very low at 0.16 which would result in 84 false positives for every 16 youths correctly identified. In general, most suicide screens are limited to adolescent and adult populations and suffer from low specificity, which may overburden programs with false positives. The benefit of being able to intervene for those true positives prior to a suicide attempt is difficult to argue; however, it is important to understand that most of those identified will be false positives. Thus, this instrument, and most suicide screening instruments in general, 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 evidence 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. Finally, issues of diversity and representativeness of the standardization sample must be considered. We revisit this issue in Chapter 8 when we discuss the future of screening research.

References

Achenbach, T. M., & Edelbrock, C. (1987). *Manual for the child behavior checklist and revised child behavior profile*. Burlington: University of Vermont Department of Psychiatry.

Ahmad, R., & Mansoor, I. (2011). What I think and feel: Translation and adaptation of revised children's manifest anxiety scale, second edition (RCMAS-2) and its reliability assessment. *The International Journal of Educational and Psychological Assessment, 8,* 1–11.

Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th edn.). Washington, DC: Author.

Ang, R. P., Lowe, P. A., & Yusof, N. (2011). An examination of the RCMAS-2 scores across gender, ethnic background, and age in a large Asian school sample. *Psychological Assessment, 23,* 899–910.

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. (2005). *Beck youth inventories* (2nd ed.). San Antonio: Psychological Corporation.

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., & 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.

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.

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. E., & Floyd, R. G. (2004). A review of the Beck youth inventories of emotional and social impairment. *Journal of School Psychology, 42*(4), 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.

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.

Caldarella, P., Young, E. L., Richardson, M. J., Young, B. J., & Young, K. R. (2008). Validation of the Systematic Screening for Behavior Disorders in middle and junior high school. *Journal of Emotional and Behavioral Disorders, 16*(2), 105–117.

Campaign for Mental Health Reform. (2005). A public health crisis: Children and adolescents with mental disorders. Congressional briefing. www.mhreform.org/kids. Accessed 1 Sept 2005.

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.

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.

Carey, M. P., Faulstich, M. E., & Carey, T. C. (1994). Assessment of anxiety in adolescents: Concurrent and factorial validities of the Trait Anxiety scale of Spielberger's State-Trait Anxiety Inventory for Children. *Psychological reports, 75,* 331–338.

Christopher, R. (2001). *Review of the multidimensional anxiety scale for children. Fourteenth mental measurements yearbook*. Lincoln: Buros Institute.

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.

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

Conners, C. K. (2008). *Conners* (3rd ed.,) Manual Multi-Health Systems. North Tonawanda, NY.

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.

Conners, C. K., Sitarenios, G., Parker, J. D., & Epstein, J. N. (1998). The revised Conners' Parent Rating Scale (CPRS-R): factor structure, reliability, and criterion validity. *Journal of abnormal child psychology, 26*(4), 257–268.

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.

Dever, B. V., Mays, K. L., Kamphaus, R. W., & Dowdy, E. (2012). The factor structure of the BASC-2 behavioral and emotional screening system teacher form, child/adolescent. *Journal of Psychoeducational Assessment, 30*(5), 488–495.

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.

Dowdy, E., Twyford, J. M., Chin, J. K., DiStefano, C. A., Kamphaus, R. W., & Mays, K. L. (2011). Factor structure of the BASC–2 behavioral and emotional screening system student form. *Psychological Assessment, 23*(2), 379.

Dowdy, E., Chin, J. K., Twyford, J. M., & Dever, B. V. (2011). A factor analytic investigation of the BASC-2 behavioral and emotional screening system parent form: psychometric properties, practical implications, and future directions. *Journal of School Psychology*, *49*, 265–280

Drummond, T. (1994). The Student Risk Screening Scale (SRSS). Grants Pass, OR: Josephine County Mental Health Program.

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.

Erford, B. T., & Hase, K. (2006). Reliability and validity of scores on the ACTeRS-2. *Measurement and Evaluation in Counseling and Development, 39*(2), 97.

Eyberg, S., & Pincus, D. (1999). *Eyberg child behavior inventory & sutter-eyberg student behavior inventory—Revised*. Psychological Assessment Resources. Odessa: 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.

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.

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.

Fraccaro, R. L., Stelnicki, A. M., & Nordstokke, D. W. (2015). Review of the Multidimensional Anxiety Scale for Children (2nd ed.). *Canadian Journal of School Psychology, 30,* 70–77.

Frick, P. J., Burns, C., & Kamphaus, R. W. (2009). *Clinical assessment of child and adolescent personality and behavior* (2nd ed.). New York: Springer.

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: 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*(7), 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.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117–135.

Goodman R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry, 38,* 581–586.

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., & 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.

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.

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

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.

Huberty, T. J. (2012). *Anxiety and depression in children and adolescents: Assessment, intervention, and prevention*. New York: Springer.

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.

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.

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. (2007). *Behavior assessment system for children,* (BASC-2): Behavioral and emotional screening system (BESS) (2nd ed.). Bloomington: Pearson

Kavan, M. G. (1992). *Review of the children's depression inventory. Eleventh mental measurements yearbook*. Lincoln: Buros Institute.

Kilgus, S. P., Chafouleas, S. M., & Riley-Tillman, T. C. (2013). Development and initial validation of the social and academic behavior risk screener for elementary grades. *School Psychology Quarterly, 28*(3), 210–226.

Kilgus, S. P., Sims, W. A., von der Embse, N. P., & Riley-Tillman, T. C. (2015). Confirmation of models for interpretation and use of the social and academic behavior risk screener (SABRS). *School Psychology Quarterly*. http://psycnet.apa.org/psycinfo/2014-40590-001. Accessed 26 June 2015.

Knoff, H. M. (1992). *Review of the children's depression inventory. Eleventh mental measurements yearbook*. Lincoln: Buros Institute.

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

Kovacs, M. (2010). *Children's depression inventory-2*. North Tonawanda: 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.

Lane, K. L., Parks, R. J., Kalberg, J. R., & Carter, E. W. (2007). Systematic screening at the middle school level: Score reliability and validity of the student risk screening scale. *Journal of Emotional and Behavioral Disorders, 15*(4), 209–222.

Lane, K. L., Kalberg, J. R., Parks, R. J., & Carter, E. W. (2008). Student risk screening scale initial evidence for score reliability and validity at the high school level. *Journal of Emotional and Behavioral Disorders, 16*(3), 178–190.

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J. H., Weisenbach, J. L., et al. (2009). A comparison of systematic screening tools for emotional and behavioral disorders: How do they compare? *Journal of Emotional and Behavioral Disorders, 17,* 93–105.

Lane, K. L., Bruhn, A. L., Eisner, S. L., & Kalberg, J. R. (2010). Score reliability and validity of the student risk screening scale: A psychometrically sound, feasible tool for use in urban middle schools. *Journal of Emotional and Behavioral Disorders, 18,* 211–224.

Lane, K., Menzies, H., Oakes, W., & Kalberg, J. (2012a). *Systematic screenings of behavior to support instruction: From preschool to high school*. New York: Guilford Press.

Lane, K. L., Oakes, W. P., Harris, P. J., Menzies, H. M., Cox, M., & Lambert, W. (2012b). Initial evidence for the reliability and validity of the student risk screening scale for internalizing and externalizing behaviors at the elementary level. *Behavioral Disorders, 37,* 99–122.

Lane, K. L., Oakes, W. P., Ennis, R. P., Cox, M. L., Schatschneider, C., & Lambert, W. (2013). Additional evidence for the reliability and validity of the student risk screening scale at the high

school level: A replication and extension. *Journal of Emotional and Behavioral Disorders, 21,* 97–115.

Lane, K. L., Oakes, W. P., Menzies, H. M., Major, R., Allegra, L., Powers, L., & Schatschneider, C. (2015). The student risk screening scale for early childhood: An initial validation study. *Topics in Early Childhood Special Education, 34*(4), 234–249.

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.

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. (2013). *Multidimensional anxiety scale for children* ((MASC-2) 2nd ed.) North Tonawanda: Multi-Health Systems.

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, J. S., Sullivan, K., & Parker, J. (1999). Test-retest reliability of the Multidimensional Anxiety Scale for Children. *Journal of anxiety disorders, 13*(4), 349–358.

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: Educational Services.

McDermott, P., Marston, N., & Stott, D. (1994). *McDermott adjustment scales for children and adolescents*. Phoenix: Ed. And Psych Associates.

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: Buros Institute.

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

Menzies, H. M., & Lane, K. L. (2012). Validity of the student risk screening scale: Evidence of predictive validity in a diverse, suburban elementary setting. *Journal of Emotional and Behavioral Disorders, 20*(2), 82–91.

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.

Oakes, W. P., Wilder, K. S., Lane, K. L., Powers, L., Yokoyama, L. T., O'Hare, M. E., & Jenkins, A. B. (2010). Psychometric properties of the student risk screening scale: An effective tool for use in diverse urban elementary schools. *Assessment for Effective Intervention, 35*(4), 231–239.

Osman, A., Gutierrez, P. M., Bagge, C. L., Fang, Q., & Emmerich, A. (2010). Reynolds adolescent depression scale—second edition: A reliable and useful instrument. *Journal of Clinical Psychology, 66*(12), 1324–1345.

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.

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.

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

Reynolds, W. M. (1989). *Reynolds child depression scale*. Lutz: Psychological Assessment Resources, Inc.

Reynolds W. M. (2002). *Reynolds adolescent depression scale* (2nd edn.). Lutz: Psychological Assessment Resources, Inc.

Reynolds, W. M. (2010). *Reynolds child depression scale* (2nd edn.). Lutz: Psychological Assessment Resources, Inc.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior assessment system for children (BASC-2)* (2 nd edn.). Circle Pines: Pearson.

Reynolds, C. R., & Richmond, B. O. (1985). *"What I think and feel" reynolds child manifest anxiety scale (RCMAS)*. Los Angeles: Western Psychological Services.

Reynolds, C. R., & Richmond, B. O. (2008). *Reynolds child manifest anxiety scale (RCMAS-2)* (2nd edn.) Los Angeles: Western Psychological Services.

Rich, B. A., & Eyberg, S. M. (2001). Accuracy of assessment: The discriminative and predictive power of the eyberg child behavior inventory. *Ambulatory Child Health, 7*(3–4), 249–257.

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 W. Damon (Series Ed.) & N. Eisenberg (Ed.), *Handbook of child psychology: Vol. 3, Social, emotional, and personality development* (5th edn., 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.

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 suicide screen: Validity and reliability of a screen for youth suicide and depression. *Journal of the American Academy of Child and Adolescent Psychiatry, 41,* 71–79.

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.

Southam-Gerow, M. A., Flannery-Schroeder, E. C., & Kendall, P. C. (2003). A psychometric evaluation of the parent report form of the State-Trait Anxiety Inventory for Children—Trait Version. *Journal of anxiety disorders, 17*(4), 427–446.

Spielberger, C. D. (1973). *State trait anxiety inventory for children*. Palo Alto: Consulting Psychological Press.

Stockings, E., Degenhardt, L., Lee, Y. Y., Mihalopoulos, C., Liu, A., Hobbs, M., & Patton, G. (2015). Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of Affective Disorders, 174,* 447–463.

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.

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.

Ullman, R. K., Sleator, E. K., & Sprague, R. L. (1988). *ADD-H comprehensive teacher's rating scale (ACTeRS)*. Champaign: MetriTech, Inc.

Ullman, R. K., Sleator, E. K., & Sprague, R. L. (1997). *ADD-H comprehensive teacher's rating scale (ACTeRS)*. Champaign: MetriTech, Inc.

Ullman, R. K., Sleator, E. K., & Sprague, R. L. (2000). ACTeRS Teacher & Parent Forms Manual. Champaign, IL.

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.

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

Walker, H. M., & Severson, H. H. (1992). Systematic screening for behavior disorders (SSBD). Longmont: Sopris West.

Walker, H. M., Severson, H. H., Todis, B. J., Block-Pedego, A. E., Williams, G. J., Haring, N. G., & Barckley, M. (1990). Systematic Screening for Behavior Disorders (SSBD) Further Validation, Replication, and Normative Data. Remedial and Special Education, 11(2), 32–46.

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.

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.

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: Buros Institute.

White, J., Connelly, G., Thompson, L., & Wilson, P. (2013). Assessing wellbeing at school entry using the strengths and difficulties questionnaire: Professional perspectives. *Educational Research, 55*(1), 87–98.