2016

Diagnostic Utility of the Social Skills Improvement System Performance Screening Guide (SSIS-PSG)

S. Kathleen Krach, Michael P. McCreery, Ye Wang, Houra Mohammadiamin and Christen K. Cirks

Diagnostic Utility of the Social Skills Improvement System Performance Screening Guide (SSIS-PSG)

S. Kathleen Krach, Ph.D.

Florida State University

Michael P. McCreery, Ph.D.

University of Nevada Las Vegas

Ye Wang

Houra Mohammadiamin

Christen K. Cirks

Florida State University

Keywords: PBS, Screening, Predictive Validity, Psychometrics, Diagnostic Utility

Abstract

Researchers investigated the diagnostic utility of the Social Skills Improvement System: Performance Screening Guide (SSIS-PSG; Elliot & Gresham, 2008). Correlational, regression, ROC, and conditional probability analyses were run to compare ratings on the SSIS-PSG subscales of Prosocial Behavior, Reading Skills, and Math Skills, to report card grades for conduct, reading, and math respectively. Respective subscales were all statistically significantly correlated with one another. In addition, all regressions indicated significant predictions for the SSIS-PSG to respective report card grades. ROC analyses for SSIS-PSG Math with math grades and SSIS-PSG Reading with reading grades were statistically significant and described as fair (Compton et al., 2010). ROC analysis for SSIS-PSG Prosocial Behavior with conduct grades was not significant and described as poor (Compton et al., 2010). In a conditional probability analysis, the variable of concern for screeners concerns false negative ratios (Compton et al., 2010); all estimates for this fell within the targeted range.

Diagnostic Utility of the Social Skills Improvement System Performance Screening Guide (SSIS-PSG)

Many schools provide children with a multi-tiered system of services for both prevention and intervention (Lane, Kalberg, & Menzies, 2009). Often, this system is represented with three tiers. Tier 1 is for all children in a given school, tier 2 is for roughly 15% of children, who need more support, and tier 3 is for about 5% of students, who need more intensive help (Basham, Israel, Graden, Poth & Wintson, 2010). This multi-tiered system goes by different names depending on the purpose. If a child is having academic problems, the system is often referred to as Response to Intervention (RTI), whereas the same framework used for children with social, emotional, and behavioral problems is referred to as Positive Behavioral Intervention Supports (PBIS) or Positive Behavioral Supports (PBS; Lane et al., 2009).

In general, these multi-tiered prevention / intervention systems have been considered highly effective (OSEP, 2009). However, this effectiveness depends on schools ensuring that the central goals of the system are met. These goals, as defined by Sugai and Horner (2006), state that the system should focus on prevention, choose evidence-based practices, and have a support system in place. To help meet these goals, each tier should incorporate the following: 1) a method for collecting and managing data, 2) knowledge of and access to research-based interventions, and 3) a system for making data-based decisions (Krach & McCreery, 2015).

Given the need to both gather and utilize data, it is vital that the chosen collection tools provide the most accurate information available. Thus, psychometrically sound data must be used to determine which children need more directed services (e.g., moved from Tier 1 to Tier 2, or moved from Tier 2 to Tier 3). The purpose of this study is to examine the diagnostic utility of

one of these instruments, the Social Skills Improvement System: Performance Screening Guide (SSIS-PSG; Elliot & Gresham, 2008), to validate its use within an RTI / PBS/ PBIS framework.

Screeners in Multi-Tiered Systems

Lane, Oaks, and Menzies (2010) write about the use of screening tools to identify at-risk students within RTI and PBIS models. They suggest that, although academic screening tools (e.g., Dynamic Indicators of Basic Early Literacy Skills [DIBELS], Good & Kaminski, 2002; AIMSweb, Pearson Education, 2008) are widely used within an RTI model, behavioral screening tools (e.g. Student Risk Screening Scale, Drummond, Eddy, Reid, & Bank, 1994; Systematic Screener for Behavior Disorders, Walker & Severson, 1992) are not as commonly utilized. Instead, schools often use office disciplinary referrals (ODRs) as their metric to determine whether students need special services (Lane et al., 2010).

ODRs can provide valuable screening information for a limited number of children. Findings indicate that about half (40% in middle school and 59% in elementary school) of all ODRs in a school are accounted for by only 5%-8% of the students (Loeber & Farrington, 1998; Sprague & Walker, 2005; Sugai, Sprague, Horner, & Walker, 2000). However, these numbers only tend to emphasize the students who are the worst offenders, and may miss children who have only moderate problems. For example, ODRs can effectively identify children with significant acting-out behavior, but they often fail to identify children with milder externalizing problems (e.g., noncompliance) or children with internalizing problems (e.g., social withdrawal) (Lane, Wehby, Robertson, & Rogers, 2007). In addition, given the non-systematic nature of ODR data, a child referred by one teacher to the office may not be referred by a different one, depending on teacher variables such as teacher tolerance for misbehavior (Pyhälto, Pietarinen, & Soini, 2015), teacher burn-out (Pas, Bradshaw, Hershfeldt, & Leaf, 2010), and amount of training (Irvin, Tobin,

Sprague, Sugai, & Vincent, 2004). Who is selected for an ODR can also be influenced by child variables such as race and ethnicity (McCarthy & Hoge, 1987; Skiba, Michael, Nardo, & Peterson, 2002) or by systems variables such as administrators offering rewards to teachers for fewer ODRs (Kern & Manz, 2004) or teachers experiencing positive behavioral change in students who receive ODRs (Lane et al., 2010). Given these inconsistencies, Lane, Parks, Kalberg, and Carter (2007) describe ODRs as having problems with interrater reliability and recommend other screeners to detect social, emotional, and behavioral problems within multi-tiered systems such as RTI and PBS/PBIS (Lane et al., 2010).

Thus, it is important for researchers to investigate other techniques that may supplement or improve upon the ODR approach. Lane and colleagues (2010; 2011) recommend several formal screening instruments for use with K-12 students. Table 1 provides specific information about each instrument listed in Lane and colleagues (2010, 2011), but expands the information to include psychometric data, subscales, and the number of items, ages / grades, and cost. Given the school-based nature of this study, only the teacher-forms of instruments are included. Also, given the sample used in this study, only instruments using U.S. or English-language are included.

Of the tests listed in Table 1, three stand out as particularly problematic for use as universal screeners. Although all three appear psychometrically sound, the Behavioral and Emotional Rating Scale: Teacher rating (BERS; Epstein & Sharma, 1998), the Strengths and Difficulties Questionnaire; Parent / Teacher Single-Sided version in U.S. English (SDQ; Goodman, 1997), and the BASC-2 Behavioral and Emotional Screening Systems: Teacher Form (BESS; Kamphaus & Reynolds, 2007) all require significant time to complete. For example, the SDQ takes about 5 minutes per child to complete (Goodman & Scott, 1999); therefore, if the average elementary-school teacher has 20 children in the classroom (U.S. NCES, 2007-2008),

the SDQ would take about 1.66 hours to complete for an entire class. The makers of the BASC-2: BESS (Kamphaus & Reynolds, 2007) clock it as taking a bit more time (5 to 10 minutes per child) thus requiring between 1.66 hours and 3.33 hours for a class of 20 students. The time requirement for the BERS is listed at about 10 minutes per child (http://www.proedinc.com) or 3.33 hours for a classroom with 20 children. Considering that gathering universal screening data may be requested about three times a year (Lane et al., 2010), the simple act of completing these questionnaires would take the equivalent of a teacher's entire school day each year.

The additional three assessments listed by Lane and colleagues (2010; 2011) are much stronger contenders for use as universal screeners. The first one, the Student Risk Screening Scale (SRSS; Drummond et al., 1994) has only seven items making it easier for teachers to complete on all children. In addition, it is free to use. The reliability data available on the SRSS indicate that the scores derived from the instrument should be considered somewhat reliable (Drummon et al., 1994; Lane, Kalberg, Parks & Carter, 2008). Lane and colleagues (2008) provided reliability data from a sample of scores from high school students on the SRSS. Specifically, they calculated internal consistency Cronbach's α values, which all exceeded .78; test-retest correlation coefficients ranging from .22 to .73, and statistically significant interrater reliability data. From the same study, convergent validity data between the SSRS and the SDQ indicated a correlation coefficient in the low to moderate range ($r = .47$) across total scores. Additional reliability and validity data for the SRSS is available in Table 1 from the manual (Drummond et al., 1994).

There are a few negative issues to note regarding the SSRS. First, the subscales all focus only on negative behaviors and provide no data on positive skills. Given the focus on positive supports / skills embedded in both the PBS and the RTI frameworks (Lane et al., 2009; OSEP,

2009), the lack of data on student strengths may make the instrument less effective within these models. Another problem is that reliability and validity data are only provided for middle- and high-school populations (Lane et al., 2007; Lane et al, 2008). The data from these studies are highly supportive of its use, but more data are suggested before using it with an elementary-aged population.

Another instrument of note is the Systematic Screening for Behavior Disorders Universal Screening (SSBD; Walker & Severson, 1992). The SSBD provides a multi-tiered method of data collection that matches that found within the PBIS model. The tier 1 (stage 1) assessment is a universal screener asking teachers to rank-order children for internalizing and externalizing problems (Lane et al., 2010). Children identified as high on one (or both) of these scales are then evaluated using more comprehensive assessment tools, and services are put into place based on these results. Unlike the SRSS, the SSBD does have an associated cost, but the expense is not prohibitive. Validity and reliability data are both sound (Walker et al, 1990; Walker, Severson, & Feil, 2010), but additional diagnostic utility data on the SSBD stage-one data collection may be needed (Menzies & Lane, 2012).

Finally, the SSIS: PGS is mentioned in Lane et al. (2011), but is not included in the list by Lane et al. (2010). This may be because of how recently the instrument was released. Although the SSIS-PSG is fairly new (Gresham & Elliot, 2008), it was derived from a more venerable and well-researched assessment tool, the Social Skills Rating System (SSRS; Gresham & Elliott, 1990). The SSIS-PSG is short (only 4 items) and targeted to evaluate academics (Math Skills and Reading Skills) as well as positive behaviors (Prosocial Behavior and Motivation to Learn). Teachers can evaluate an entire class using one sheet of paper, making it a quick process. The following section discusses specific psychometric data about the SSIS-PSG.

Psychometric Data from the SSIS Manual

The authors of the SSIS-PSG provide respectable data to support the reliability of the scores on the SSIS-PSG (Gresham & Elliot, 2008). Because each subscale only has one question, no internal consistency data were available. Instead, they conducted a test-retest reliability study using 25 teachers and 543 students, with an average time between administrations of around 74 days. For the preschool-aged population, test-retest intraclass reliability coefficients ranged from 0.53 to 0.62. For K-12 students, test-retest reliability coefficients ranged from .56 to .74.

The authors also conducted an interrater-reliability assessment using 44 teachers across 434 students. Interrater reliability was derived by comparing scores from the child's main teacher to another individual who was familiar to the child and was nominated by the main teacher. The interrater, intraclass reliability coefficient for preschoolers ranged from .60 to .73, and for elementary students, ranged from .55 to .68. It is at the secondary grades that the interrater, intraclass reliability decreased to a range of between .37 (Prosocial Behavior) and .60. The authors cite standards set by Landis and Koch (1977) as an indication that these data demonstrate moderate score reliability. The biggest problem with using these Landis and Koch (1977) standards is that they are meant to be used with categorical data and not interval data. So, depending on how a person views rating scale data (e.g., categorical, interval, etc.), the standards used to determine reliability might differ, as might the statistical analyses used to evaluate these standards (Jamieson, 2004).

It is in the validity section that the SSIS-PSG manual begins to show weakness. First, the SSIS-PSG manual does not provide direct information about the content validity of the smaller instrument (PSG) but only of the SSIS (full version). This is particularly problematic, as the

items on the PSG were not drawn from the larger SSIS assessment; the authors provide no information as to how the items were selected. Therefore, the content validity of the larger SSIS does not apply, and without content validity information, it is difficult to evaluate the development of the items in this tool.

The only validity data for the PSG were in the form of concurrent validity; the authors compared the SSIS Performance Screening Guide to the teacher form of the larger instrument (SSIS Rating Scales). The SSIS-PSG authors could have compared it to any of the other choices in Table 1, but instead compared it to their own instrument. There may be criterion-contamination issues when comparing one's own instruments to each other for validation (Dowling, 1986).

In addition, although the SSIS-PSG measures four areas (Prosocial Behavior, Motivation to Learn, Reading Skills, and Math Skills), they compared those ratings to only two areas for preschoolers on the SSIS: Rating Scales (Social Skills and Problem Behaviors) and three scales for the elementary and secondary students (Social Skills, Problem Behaviors, and Academic Competencies). As is clear from the names of these comparison scales, they fail to provide an accurate or differentiated comparison for the areas of reading, math, and motivation to learn. Therefore, the SSIS-PSG manual significantly underrepresents the data needed to make an informed decision about the validity of this instrument for use within a RTI or PBS/PBIS framework.

Psychometric Data from the Literature

Given the paucity of information in the manual on validity for the SSIS-PSG, the current researchers conducted a review of the current literature available. Given how new the instrument is, there were a surprising number of articles to be found. Several of these simply describe the

SSIS, but provide no further data specific to the SSIS-PSG (Gresham, Elliott, Cook, Vance, & Kettler, 2010; Gresham, Elliott, Vance, & Cook, 2011). At this time, three published test reviews of the SSIS are available (but none specific to only the SSIS-PSG). Two of these (Doll & Jones, 2010; Lee-Farmer, 2010) are from the Buros Mental Measurement Yearbook. Of these two reviews, only Lee-Farmer (2010) mentions the inclusion of the SSIS-PSG in the overall instrument. This review does not evaluate the SSIS-PSG, but simply mentions its inclusion in the materials. A review of the SSIS published in the Journal of Psychoeducational Assessment (Crosby, 2011) mentions that the SSIS-PSG is a screening component of the SSIS, but does not critique it separately from the full-scale assessment.

There do exist a few studies of note about the validity of the SSIS-PSG. In the first, Lane et al. (2015) published a convergent validity study of the separate parts of the SSIS-PSG and the composite for the SRSS-IE (adapted from the original SRSS). All of the correlations were significant ($p < .0001$) and negative: Reading Skills (-0.41), Math Skills (-0.43), Motivation to Learn (-0.62), and Prosocial Behavior (-0.72). This negative relationship is expected, given that the SSIS-PSG examines variables of positive behavior and the SRSS-IE examines variables of negative behavior. In the second, Lane, Richards-Tutor, Oakes, and Connor (2013) provided concurrent validity between the SRSS and the SSIS-PSG with English Language Learners. Again, they found a negative and statistically significant ($p < .0001$) coefficient for each of the relationships: Reading Skills (-0.50), Math Skills (-0.50), Motivation to Learn (-0.63), and Prosocial Behavior (-0.53).

Finally, Kettler, Elliot, Davies, and Griffin (2012) included the only diagnostic utility study available on the SSIS-PSG. They used an Australian sample of 360 students in both 3rd and 5th grades. These authors provided information about the sensitivity of the SSIS-PSG for

identifying academic weaknesses on standardized tests. They identified a cut-off score for both the SSIS-PSG (scores of 1, 2, or 3 were indications of below expectations) and used passing scores on the national achievement test (as set by the nation of Australia) as the predictive cut-off. Agreement between the two for sensitivity (a ratio of true positives to false negatives for individuals failing to meet the standard) was 0.95, and for specificity (a ratio of false positives to true negatives for individuals at or above the standard) was 0.45. They also provided information on positive predictive value (a ratio of true positives to the total identified by the SSIS-PSGs) of 0.20 and negative predictive value (a ratio of true negatives to the total not identified by the SSIS-PSGs) of 0.99.

There were some problems of note with the Kettler and colleagues (2012) study. First, although the sample was from an English-speaking country, the data collected was not based on standards in the United States. This is especially important given that the predictor was performance on an assessment tool based off of nationally decided upon cut-off scores. In addition, the sample only consisted of two grades of students, and as is reported in the article, the score differences between these two grades were significant. This limited grade range does not provide sufficient information to state that the SSIS-PSG is predictive across elementary school. However, the most troubling issue within the article is that the authors do not describe which of the four SSIS-PSG subscales (Prosocial Behavior, Motivation to Learn, Math Skills, or Reading Skills) were used in the analysis; they identify the SSIS-PSG as a single variable (the PSGs) that may or may not be a composite. Given a hypothesized interconnected relationship between these variables (Zins, Bloodworth, Weissberg, & Walberg, 2007), a composite may be useful for consideration by school-based personnel. However, since this relationship is not absolute, a composite should not be seen as sufficient for intervention decisions or planning (Krach,

McCreery, & Vallett, 2016). More specific data about which SSIS-PSG variable predicts which aspect of academic performance is needed to ensure the validity of the instrument within an RTI/ PBS/ PBIS model.

Menzies and Lane (2012) have already pointed out this need for more validity information on screening tools used in multi-tiered systems. They specifically call for more research on the psychometrics of tools for PBS/PBIS. The authors state that, although sufficient research is being conducted on the SRSS, they "recommend other research teams explore additional systematic screening tools, including the …SSIS-PSG … to determine if these instruments hold equal (or greater) predictive accuracy in behavioral, social, and academic domains" (p. 90). Thus, the goal of the current paper is to examine the diagnostic utility of the SSIS-PSG to predict academic grades in elementary-aged children.

Method

Participants

Ten classroom teachers from a Title 1, urban school in the Southeastern United States volunteered that their classes be part of this study. Out of their 207 students, 170 had consent forms signed by their parents allowing them to be included in the study. However, nine students withdrew from school during the study and 42 had incomplete data sets so they were removed from this study. Given this, a total of 119 students were included in the final sample.

Students and teachers were fairly evenly divided across grades: first grade (two teachers, 35 students); second grade (two teachers, 20 students), third grade (one teacher, 18 students), fourth grade (two teachers, 32 students), and fifth grade (one teacher, 14 students). All of the teachers were female and African-American. All of the children (n = 119) were described as

African-American. The children's gender was almost evenly divided: male (n = 57) and female (n = 62).

Instruments

SSIS-PSG. The Social Skills Improvement System: Performance Screening Guide (SSIS-PSG; Elliot & Gresham, 2008) is a set of four single-item scales rated on a 5-point ordinal scale assessing a child's abilities in the areas of Math Skills, Reading Skills, Prosocial Behaviors, and Motivation to Learn. Teachers are asked to rate students' ability in each of these areas, with "1" representing very limited skills and "5" representing excellent skills. Additional information on this instrument is available in Table 1 and in the literature review of this paper. Teachers completed the SSIS-PSG in January of the school year on all of their children; however, only those whose parents provided consent were included in the study.

Teacher-issued report card grades. There is a debate as to which is a better indicator of school success variables: report card grades or standardized test scores (Jussim, 1991; Sattler, 2008; Wentzel, 1989). Given that neither test scores nor report card grades are perfect indicators of academic skills, report card grades were chosen for this study because they provide inclusive measures of reading and math skills and not simply single subtypes such as reading fluency or math reasoning. In addition, report card grades provide an estimate of prosocial skills through conduct grades.

Report card grades for those whose parents provided consent were obtained from the students' records. The report cards were dated November of the same school year as the SSIS-PSG ratings (within two months apart). Report card grades for reading and math were provided on a scale out of 100 total points, reflecting percentage correct on class assignments such as tests and homework. The child's conduct grade was converted from a qualitative indication of a

teacher's judgment of conduct to a 5-point ordinal rating scale: 1 = unsatisfactory, 2 = needs improvement, 3 = satisfactory, 4 = good, 5 = exceptional.

Unfortunately, there was no corresponding measure of Motivation to Learn included in this study. Therefore, this item on the SSIS-PSG was not included in the current study.

Data Analysis

First, descriptive statistics were run for each variable, including skewness and kurtosis for each of the variables (see Table 2). There are discrepancies in opinions by statisticians in this area. Therefore, acceptable limits of skewness and kurtosis may be either an absolute value of 2 (Tabachnick & Fidell, 1996) or it can be an absolute value greater than 2 but less than 3 (Richards & Gross, 2005). For the current study, anything with an absolute value of 2 was deemed sufficient.

Second, Spearman Rho correlation coefficients were calculated using SPSS version 22 for report card grades to SSIS-PSG ratings. These correlations can be found in Table 3. Third, a conditional probability analysis was run to examine how well the SSIS-PSG predicted failing cut-off points for grades. As with Kettler and colleagues (2012), scores on the SSIS-PSG of one, two, or three were rated as failing to meet the standard, while four and five were at or above the standard. Academic grades of "D" (60-69) or "F" (59 or below) were rated as failing to meet the standard. With conduct grades, any grade of "unsatisfactory: 1" or "needs improvement: 2" was deemed as not meeting the standard. Tables 5 and 6 provides this information.

The conditional probability analysis measures four things (Kettler et al., 2012, p.95): sensitivity (the likelihood that a screening test will correctly identify a student as below the minimum cut-off), specificity (the likelihood that a screening test will correctly not identify a student who is at or above the minimum cut-off), positive predictive value (PPV; the likelihood

that an identified student is below the minimum cut-off), and negative predictive value (NPV; the likelihood that a student who is not identified is at or above the minimum cut-off).

In addition, four other calculations were conducted based on descriptions by Kessel and Zimmerman (1993). False positive rates provide a percentage calculated by dividing the number of false positives by the total number of individuals who should not have been identified as having a problem. False negatives rates also provide a percentage, but this time it is calculated by dividing the number of false negatives by the total number of individuals who should have been identified as having a problem. Overall correct classification provides a percentage calculated by dividing the true positives and true negatives by the sample total. Finally, Kappa was calculated; Kappa is described by Kessel and Zimmerman as "the level of agreement between the test in question and a gold standard beyond that accounted for by chance alone" (1993, p. 395).

As part of the conditional probability analysis, data for sensitivity, specificity, positive predictive value, negative predictive value, false positive rates, false negative rates, overall correct classification, and Kappa were calculated and placed in Table 5 for corresponding variables. Specifically, report card grades in reading were compared to SSIS-PSG Reading Skills ratings, math report card grades were compared to SSIS-PSG Math Skills ratings, and conduct report card grades were compared to SSIS-PSG Prosocial Behavior ratings using the formulae described in Kessel and Zimmerman (1993).

Finally, a Receiver Operating Characteristic (ROC) curve analysis using SPSS version 22 was also calculated to determine how well the SSIS-PSG subscales predicted at-risk status based on the previously outlined cutoff scores. ROC curve analyses are intended to examine the trade-off between false positive and false negative rates of a diagnostic assessment. Area under the curve (AUC) scores of .5 or less suggest the accuracy of the test is no better than chance, while

findings of .80 and above suggest good to strong indicators of diagnostic accuracy (McFall & Treat, 1999; Swets, 1992). For the purposes of this study, scores of needs improvement and below (i.e., 2 or 1) were considered a positive test.

Results

Table 3 provides information on the correlations between the report card grades and the SSIS-PSG, while Table 4 provides regression data. Table 5 provides information on the true and false positives for the SSIS-PSG that is later used to calculate data for the conditional analysis framework indices. Table 6 provides information for each of the conditional analysis framework indices. Because these indices are fluid (if one increases the standard for false positives, one may decrease the number of false negatives), a method for determining how to interpret these findings depends on the goals for the instrument. For the SSIS-PSG, the goals should be set based upon needs for a screening tool (Compton et al., 2010). The comparison column on Table 6 provides both ideal targets for screeners as well as comparison score ranges to similar screening instruments from the literature.

Regarding sensitivity, all conditional analysis indices were both on target and within the typical range. For specificity, all conditional analysis indices were below target and below range. In the case of positive predictive value, all were below range and well below the target; whereas for negative predictive value, all were above both the range and the target. For the false positive rate, all were in range but below the target; however, for the false negative rate, all were both in range and met the target. For overall correct classification, all are below the range provided by other instruments except for Pro-Social Behavior when compared to conduct report card grades; this one was within range. None of the overall correct classification indices met the desired

target. Kappa agreement scores are very low, indicating a level of agreement between the test and an ideal version of itself.

Findings from the ROC curve analysis indicated that the SSIS-PSG Reading Skills (Figure 1) and SSIS-PSG Math Skills (Figure 2) subscale provided a fair level of discriminatory power for identifying reading (area under the curve [AUC] = .743; p = .002; CI = .627 to .858) and math ([AUC] = .747; p = .001; CI = .627 to .868) deficiencies; however, neither reached the recommended .800 threshold (McFall & Treat, 1999; Swets, 1992). Further, SSIS-PSG Prosocial Behavior (Figure 3) provided findings for discriminatory power similar that are similar to chance (AUC = .565; p = .477; CI = .400 to .731).

Discussion

Multi-tiered prevention and interventions in schools (e.g., RTI and PBS/PBIS) depend upon psychometrically sound assessment tools to determine children's placement needs for services (Krach & McCreery, 2015). When children are initially identified for additional services (moving from Tier 1 to Tier 2), good screening instruments are the preferred method of helping to make this determination (Lane et al., 2010). But what constitutes a good screener? Glover and Albers (2007) suggested that a good behavioral screener should be efficient (not require too much time) and accurate (psychometrically sound), while also displaying reasonable levels of specificity (can identify exactly what is wrong) and sensitivity (can identify all of the children who need help). In addition, all of this should exist within the most cost-effective assessment available (Lane et al., 2015).

Reading through this list of requirements, it is clear that some of these are impossibly contradictory. For example, it is often difficult for an instrument to be both brief and also reliable (Onwuegbuzie & Daniel, 2002). In addition, these brief and accurate screeners must also be

general enough to catch all children at risk (requiring them to have enough questions to cover multiple types of behaviors) while also being specific enough for intervention planning purposes (requiring them to have multiple questions about specific types of problems). Table 1 provides information about the most currently accepted and commonly used screeners for these purposes. As can be seen, these screeners range from having one subscale to seven subscales. The number of items ranges from three to 52. And, finally the costs range from free to about $1.50 per child per administration. Given the limited number of screeners, it is imperative that all be evaluated to ensure sufficient psychometric accuracy (Menzies & Lane, 2012), especially when used as part of an RTI / PBS / PBIS framework.

On the list from Table 1, it is clear that the SSIS-PSG is the one most in need of additional review. The goal of this article was to explore the diagnostic utility of the SSIS-PSG within an elementary-school population. There are many different methods of exploring diagnostic utility. One way is to explore the predictive validity of the instrument. Although no set correlation standards have been established to determine sufficient predictive validity, a few guidelines do exist. One states that, if two data sets show statistically significant correlations with one another, then the one under review may be considered a valid addition to a test battery (Anastasi & Urbina, 1988). A second guideline states that predictive validity may be evident for statistically significant correlations, even if the coefficients are “as low as .20 or .30” (Anastasi & Urbina, 1988, p. 144). For the SSIS-PSG, all three respective correlations were statistically significant at the .01 level: SSIS-PSG: Math to math grades (.52), SSIS-PSG Reading to reading grades (.53), Prosocial Behavior to conduct grades (.24). Therefore, based on the Anastasi and Urbina (1988) standards, there is evidence of predictive validity. In addition to these correlations, a regression analysis also was conducted for each of these variable combinations. All of these regressions

resulted in statistically significant findings; however, the effect size estimates for the math and reading pairings were only moderate and the behavioral pairing was low.

Although the regression and correlation data provide a clear indication of the relationships between the SSIS-PSG ratings and report card grades, the utility of the instrument needs accuracy estimates in addition to simple predictive analyses. To that end, an ROC analysis was conducted on each corresponding pair to determine the discriminatory power of the SSIS-PSG in decision making. While the discriminatory power should be considered fair for the math and reading pairings, the behavioral pairing's discriminatory power should be viewed as no better than chance.

Another measure of utility, a conditional probability analysis, was conducted using the guidelines set forth in Kessel and Zimmerman (1993) examining eight indices: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), false positive rates, false negative rates, overall correct classification rates, and Kappa. Next, the SSIS-PSG indices' data were compared either to standards set within the field or to statistical data provided for similar instruments. Based on the results of the conditional probability analysis of similar screeners along with the SSIS-PSG (presented in Table 6), it is clear that there should be some concerns about the accuracy of all of the screeners.

Given that all of the screeners examined failed to meet at least one of the eight target indices on the conditional probability analysis, it is important to establish which index should hold the most weight in instrument utility decisions. To choose the index (or indices) of most importance, one must first identify the purpose of the assessment. As the SSIS-PSG is a screening tool (not a diagnostic one), the most important index of concern is the one estimating "false negatives" (Compton et al., 2010; Landau, Milich, & Widiger, 1991). The reason that this estimate is most

valued for screeners is because a screening tool is just the first gate in a multi-tiered system for helping children in school. As such, it is less troubling when children are over-identified as needing help when they don't need it (false positives) than not getting help when they do need it (false negatives). If a screener falsely identifies a child as needing help, this child will be correctly identified at one of the later tiers of the system. However, children who do need help but are not identified for services (false negatives) may not get help until they demonstrate a serious problem. The false negative conditional probability analysis index results for all three sections analyzed in this study for the SSIS-PSG were not only similar to other screeners but also met the expected target range set by Compton and colleagues (2010).

When comparing all of these usability analyses (correlations, regression, ROC, etc.) on the SSIS-PSG to one another, a few trends appear. The SSIS-PSG seems to have sound utility in screening for both math and reading problems across most of these analyses. However, this is not true for the Prosocial Behavior scale and conduct grades analyses. These analyses find a clear relationship between the two, and the Prosocial Behavior scale does seem to accurately identify children with overall conduct problems. However, other utility analyses did not show Prosocial Behavior as a useful estimate for predicting positive conduct.

A closer look at these issues may find that the problem lies with construct differences between Prosocial Behavior and conduct grades. Instead of measuring a positive construct such as prosocial skills, conduct grades may instead measure the absence of a negative set of behaviors (authors, in submission). This difference is subtle, but it may be the overall cause for the conflicting nature of some of the analyses in this study. For example, if the goal of the screener is to identify children who need help increasing positive skills to substitute for negative behaviors, then it seems that the Prosocial Behavior scale on the SSIS-PSG may do a fine job.

However, it may be impossible to tell from the data in the current study if the SSIS-PSG would also be able to identify children who already have good prosocial behaviors and need no additional support.

There are a few other issues in this study that could be improved upon. The first is that, although report card grades may be good indicators of academic ability, their use as diagnostic indicators may be problematic (Jussim; 1991; Wentzel, 1989). Instead, it may benefit future researchers to examine the SSIS-PSG against both report card grades and standardized test scores to ensure maximum accuracy. Another concern with this study is that certain age groups were not included (Pre-K, K, middle, and high school). Therefore, additional research on diagnostic utility for these populations needs to be considered.

It is clear from the current study and the Kettler et al. (2012) study that the SSIS-PSG provides some predictive information about student success. But, as is true of the use of office disciplinary referrals (ODRs; Lane, Oaks, & Menzies, 2010), it is also clear that the SSIS-PSG should not be a stand-alone method of making PBIS tier placement decisions. Just as with any placement decision, accurate assessment most likely comes from a combination of multiple methods and/or multiple raters for complete accuracy (Sattler, 2008).

References

Anastasi, A., & Urbina, S. (1988). Psychological testing: Seventh edition. Upper Saddle River, NJ: Prentice Hall.

Basham, J. D., Israel, M., Graden, J., Poth, R., & Winston, M. (2010). A comprehensive approach to RTI: Embedding universal design for learning and technology. Learning Disability Quarterly, 33(4), 243-255. doi: 10.1177/073194871003300403

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression / correlation analysis for the behavioral sciences. Mahwah, NJ: Erlbaum.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., ... & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. Journal of Educational Psychology, 102(2), 327. doi: 10.1037/a0018448

Crosby, J. W. (2011). Test Review: F M Gresham & S N Elliott "Social Skills Improvement System Rating Scales." Minneapolis, Minnesota--NCS Pearson, 2008. Journal of Psychoeducational Assessment, 29(3), 292-296. Doi: 10.1177/0734282910385806

Distefano, C., & Kamphaus, R. W. (2007). Development and validation of a behavioral screener for preschool-aged children. Journal of Behavioral and Emotional Disorders, 15, 93-102. doi: 10.1177/10634266070150020401

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. doi: 10.1177/0734282912438869

Doll, B., & Jones, K. (2010). Review of the Social Skills Improvement System Rating Scales. In R. A. Spies, J. F. Carlson, K. F. Geisinger, & L. L. Murphy (Eds.), Mental Measurement Yearbook (pp. 561-565). Lincoln, NE: University of Nebraska Buros Institute of Mental Measurements.

Dowling, G. R. (1986). Perceived risk: The concept and its measurement. Psychology & Marketing, 3, 193-210.

Drummond, T., Eddy, J. M., Reid, J. B., & Bank, L. (1994, November). The Student Risk Screening Scale: A brief teacher screening instrument for conduct disorder. Paper presented at the 4th annual Prevention Conference, Washington, DC.

Elliot, S. N., & Gresham, F. M. (2008). Social Skills Improvement System: Classwide *intervention program teacher's gu*ide. Bloomington, MN: Pearson / PsychCorp.

Epstein, M. H., Harniss, M. K., Pearson, N., & Ryser, G. (1999). The Behavioral and Emotional Rating Scale: Test-retest and inter-rater reliability. Journal of Child and Family Studies, 8(3), 319-327. doi: 10.1023/A:1022067329751

Epstein, M. H., & Sharma, H. M. (1998). Behavioral and Emotional Rating Scale. Austin, TX: Pro-Ed.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. Journal of School Psychology, 45(2), 117-135. doi:10.1016/j.jsp.2006.05.005

Good, R. H., Kaminski, R. A., & Dill, S. (2002). Dynamic Indicators of Basic Early Literacy Skills, 6 (DIBELS). Eugene, OR: University of Oregon Center on Teaching and Learning

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. Journal of Child Psychology and Psychiatry, 38, 581–586. doi: 10.1111/j.1469-7610.1997.tb01545

Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is small beautiful? Journal of Abnormal Child Psychology, 27(1), 17-24. doi: 10.1023/A:1022658222914

Gredler, G. R. (1997). Issues in early childhood screening and assessment. Psychology in the Schools, 34, 99–106. doi: 10.1002/(SICI)1520-6807(199704)

Gresham, F. M., & Elliott, S. N. (1990). Social Skills Rating System. Minneapolis, MN: Pearson Assessments.

Gresham, F. M., & Elliott, S. N. (2008). Social Skills Improvement System-Rating Scales. Minneapolis, MN: Pearson Assessments.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System—Rating Scales. Psychological Assessment, 22(1), 157. doi: 10.1037/a0018124

Gresham, F. M., Elliott, S. N., Vance, M. J., & Cook, C. R. (2011). Comparability of the Social Skills Rating System to the Social Skills Improvement System: Content and psychometric comparisons across elementary and secondary age levels. School Psychology Quarterly, 26, 27. doi: 10.1037/a0022662

Guilford, J. P., & Fruchter, B. (1978). Fundamental statistics in psychology and education: Sixth edition. New York, NY: McGraw-Hill Book Company.

Harniss, M. K., Epstein, M. H., Ryser, G., & Pearson, N. (1999). The Behavioral and Emotional Rating Scale convergent validity. Journal of Psychoeducational Assessment, 17, 4-14. doi: 10.1177/073428299901700101

Irvin, L. K., Tobin, T. J., Sprague, J. R., Sugai, G., & Vincent, C. G. (2004). Validity of office discipline referral measures as indices of school-wide behavioral status and effects of school-wide behavioral interventions. Journal of Positive Behavior Interventions, 6(3), 131-147. doi: 10.1177/10983007040060030201

Jamieson, S. (2004). Likert scales: How to (ab)use them. Medical Education, 38(12), 1217-1218. doi: 10.1111/j.1365-2929.2004.02012

Jussim, L. (1991). Grades may reflect more than performance: Comment on Wentzel (1989). Journal of Educational Psychology, 83 (1), 153-155. doi: 10.1037/0022-0663.83.1.153

Kamphaus, R. W., & Reynolds, C. R. (2007). BASC-2 Behavioral and Emotional Screening System (BASC-2: BESS). Circle Pines, MN: AGS Publishing.

Kern, L., & Manz, P. (2004). A look at current validity issues of school-wide behavior support. Behavioral Disorders, 30, 47-59.

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self–administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. Psychological Assessment, 5, 395–399. doi:10.1037/1040-3590.5.4.395

Kettler, R. J., Elliott, S. N., Davies, M., & Griffin, P. (2012). Testing a multi-stage screening system: Predicting performance on Australia's national achievement test using teachers' ratings of academic and social behaviors. School Psychology International, 33(1), 93-111. doi: 10.1177/0143034311403036

Krach, S. K., & McCreery, M. P. (2015). Technology and positive behavioral support: Evaluation, selection, and implementation of computer-based socio-emotional training in schools. In Tettegah, S. Y. & Espelage, D. (Eds.), Emotions and technology: Communication of feelings for, with, and through digital media – Volume I: Emotions, Learning, and Technology. Waltham, MA: Elsevier.

Krach, S. K., McCreery, M. P., & Vallett, D. (2016). Interconnected Relationships: Academic Skills, School Behaviors, and Motivation to Learn. Presentation at the National Association of School Psychologists Annual Convention, New Orleans, LA.

Landis, J. R., & Koch, G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174. doi: 10.2307/2529310

Landau, S., Milich, R., & Widiger, T. A. (1991). Predictive power methods may be more helpful in making a diagnosis than sensitivity and specificity. Journal of Child and Adolescent Psychopharmacology, 1, 343-351. doi:10.1089/cap.1991.1.343

Lane, K. L., Kalberg, J. R., & Menzies, H. M. (2009). Developing schoolwide programs to prevent and manage problem behaviors: A step-by-step approach. New York: Guilford.

Lane, K. L., Kalberg, J. R., Menzies, H., Bruhn, A., Eisner, S., & Crnobori, M. (2011). Using systematic screening data to assess risk and identify students for target supports: Illustrations across the K-12 continuum. Remedial and Special Education, 32, 39-54. doi: 10.1177/0741932510361263

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, 178-190. doi: 10.1177/1063426608314218

Lane, K. L., Oakes, W. P., Common, E. A., Zorigian, K., Brunsting, N. C., & Schatschneider, C. (2015). A comparison between SRSS-IE and SSIS-PSG scores: Examining convergent validity. Assessment for Effective Intervention, 40(2), 114-126. doi:10.1177/1534508414560346

Lane, K. L., Oakes, W., & Menzies, H. (2010). Systematic screenings to prevent the development of learning and behavior problems: Considerations for practitioners, researchers, and policy makers. Journal of Disability Policy Studies, 21(3), 160-172. doi: 10.1177/1044207310379123

Lane, K. L., Parks, R. J., Kalber, 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, 209-222. doi: 10.1177/10634266070150040301

Lane, K. L., Richards-Tutor, C., Oakes, W. P., & Connor, K. (2013). Initial evidence for the reliability and validity of the student risk screening scale with elementary age English learners. Assessment for Effective Intervention, 39, 2190232. doi: 10.1177/153450413496836

Lane, K. L., Wehby, J., Robertson, E. J., & Rogers, L. (2007). How do different types of high school students respond to positive behavior support programs? Characteristics and responsiveness of teacher-identified students. Journal of Emotional and Behavioral Disorders, 15, 3–20. doi: 10.1177/10634266070150010201

Lee-Farmer, J. (2010). Review of the Social Skills Improvement System Rating Scales. In R. A. Spies, J. F. Carlson, K. F. Geisinger, & L. L. Murphy (Eds.), Mental Measurement Yearbook (pp. 565-566). Lincoln, NE: University of Nebraska Buros Institute of Mental Measurements.

Loeber, R., & Farrington, D. P. (Eds.). (1998). Serious and violent juvenile offenders: Risk factors and successful interventions. Thousand Oaks, CA: Sage.

McCarthy, J. D., & Hoge, D. R. (1987). The social construction of school punishment: Racial disadvantage out of universalistic process. Social Forces, 65, 1101–1120. doi: 10.1093/sf/65.4.1101

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

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. doi: 10.1177/1063426610389613

Milich, R., Widiger, T. A., & Landau, D. (1987). Differential diagnosis of attention deficit and conduct disorders using conditional probabilities. Journal of Consulting and Clinical Psychology, 55, 762-767. doi:10.1037/0022-006X.55.5.762

Office of Special Education Programs (OSEP; March 2009). Is School-Wide Positive Behavior Support an Evidence-Based Practice? Retrieved from http://www.pbis.org/research

Onwuegbuzie, A. J., & Daniel, L. G. (2002). A framework for reporting and interpreting internal consistency reliability estimates. Measurement and Evaluation in Counseling and Development, 35(2), 89.

Pas, E. T., Bradshaw, C. P., Hershfeldt, P. A., & Leaf, P. J. (2010). A multilevel exploration of the influence of teacher efficacy and burnout on response to student problem behavior and school-based service use. School Psychology Quarterly, 25(1), 13. doi: 10.1037/a0018576

Pearson Education (2014). AIMSWeb. New York, NY: NCSPeasron, Inc.

Pyhältö, K., Pietarinen, J., & Soini, T. (2015). When teaching gets tough–Professional community inhibitors of teacher-targeted bullying and turnover intentions. Improving Schools, 1365480215589663. doi: 10.1177/1365480215589663

Reynolds, C. R., & Kamphaus, R. W. (2004). Behavior Assessment System for Children (2nd ed.). Circle Pines, MN: AGS Publishing.

Richards, J. M., & Gross, J. J. (2006). Personality and emotional memory: How regulating emotion impairs memory. Journal of Research in Personality, 40(5), 631-651.

Sattler, J.M. (2008). Assessment of children: Cognitive applications (5th ed.). San Diego, CA: Author.

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. The Urban Review, 34(4), 317-342. doi: 10.1023/A:1021320817372

Sprague, J. R., & Walker, H. M. (2005). Safe and healthy schools: Practical prevention strategies. Guilford Press.

Sugai, G., & Horner, R. R. (2006). A promising approach for expanding and sustaining school-wide positive behavior support. School Psychology Review, 35, 245-259.

Sugai, G., Sprague, J. R., Horner, R. H., & Walker, H. M. (2000). Preventing School Violence: The use of office discipline referrals to assess and monitor school-wide discipline Interventions. Journal of Emotional and Behavioral Disorders, 8(2), 94-101. doi: 10.1177/106342660000800205

Swets J.A. (1992). The science of choosing the right decision threshold in high-stake diagnostics. American Psychologist, 47, 522–532. doi: 10.1037/0003-066X.47.4.522

Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics (3rd edition). New York: Harper Collins.

U.S. Department of Education, National Center for Education Statistics (NCES, 2007-2008). Schools and Staffing Survey (SASS): Public school teacher data file. Retrieved at http://nces.ed.gov/surveys/SASS/tables/sass0708_2009324_t1s_08.asp

Walker, H. M., & Severson, H. (1992). Systematic Screening for Behavior Disorders: Technical manual. Longmont, CO: Sopris West.

Walker, H. M., Severson, H. H., & Feil, E. G. (2010). Systematic Screening for Behavior Disorders (SSBD). 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. doi: 10.1177/074193259001100206

Wentzel, K. R. (1989). Adolescent classroom goals, standards for performance, and academic achievement: An interactionist perspective. Journal of Educational Psychology, 81, 131–142. doi: 10.1037/0022-0663.81.2.131

Zins, J. E., Bloodworth, M. R., Weissberg, R. P., & Walberg, H. J. (2007). The scientific base linking social and emotional learning to school success. Journal of Educational and Psychological Consultation, 17(2-3), 191-210. doi: 10.1080/10474410701413145

Table 1

Comparison of Popular PBS Screening Tools

Instruments	Subscales	# of Items	Ages / Grades	Cost	Reliability Data	Validity Data
Behavioral and Emotional Rating Scale: Teacher Rating (BERS; Epstein & Sharma, 1998)	Interpersonal Strengths Family Involvement Intrapersonal Strengths School Functioning Affective Strengths	52	Ages 5-0 to 18-11	$198 for kit $37 (pack of 25 children)	Internal Consistency >.80 - > .90 Test-Retest Alternative High School: .86 - .99 EBD High School: .83-.98	Content Validity Clearly outline in manual Construct Validity 5 Clear Factors Convergent Validity BERS Total with Walker-McConnell Total.77
Social Skills Improvement System: Performance Screening Guide (Gresham & Elliot, 2008) SSIS-PSG to SRSS-IE study by Lane et al. (2015)	Math Skills Reading Skills Motivation to Learn Prosocial Behavior	4	Grades Pre-K Grades K-6 Grades 7-12	$45.45 (pack of 10 classes)	Test-Retest: Pre-K: 0.53-0.62 K-12: .56 - .74 Interrater: Pre-K: .60-.73 K-6: .55 - .68 7-12: .37-.60	Concurrent Validity: SSIS-PSG to SSIS Rating Scales Ages 5-18 Prosocial to Social Skills .70; Academic Competence to Reading .67 and Math .60 Prosocial - Social Skills .70 Concurrent Validity SSIS-PSG to SRSS-IE Read -.41; Math -.43; MtL -.52, Prosocial -.72
Strengths and Difficulties Questionnaire; Parent / Teacher Single-Sided version in U.S. English (SDQ; Goodman, 1997)	Emotional Problems Conduct Problems Hyperactivity Peer Problems Prosocial	25	Ages 2-4 4-10 11-17	Free	Interrater: Ages 4-16: .37 - .62	Divergent Validity Parent .87; teacher .85 Concurrent Validity: SDQ with Rutter Parent .88; Teacher .92 SDQ with CBCL .59 - .87

Table 1

Comparison of Popular PBS Screening Tools (cont.)

Instruments	Subscales	# of Items	Ages / Grades	Cost	Reliability Data	Validity Data
Student Risk Screening Scale (SRSS; Drummond, 1994)	Steal Lie, Cheat, Sneak Behavior Problems Peer Rejection Low Academic Achievement Negative Attitude Aggressive Behavior	7	Grades K-12	Free	Internal Consistency High School: .78 - .86 Middle School: .78 - .85 Test-Retest: High School: .22-.71 Middle School: .56-.83 Interrater: High School: .22-.73	Convergent Validity: SRSs to SDQ High School Total to Total -.47 Middle School Total to Total .66
Systematic Screening for Behavior Disorders Universal Screening (SSBD; Walker & Severson, 1992)	Internalizing Externalizing Normal	3	Grades PreK-9	$30 (pack of 100 children)	Interrater: Externalizing .89 - .94 Internalizing .82-.90 Test – Retest Externalizing: .81-.88 Internalizing .74-.79	Discriminate Valdiity 68% to 95% Stage 1 screening later identified. 90% of EBD children ranked highly on Externalizing
BASC-2 Behavioral and Emotional Screening Systems: Teacher Form (BESS; Kamphaus & Reynolds, 2007)	F Index (Validity) Consistency (Validity) Response Pattern (Validity) Form Scores (Total)	25-30	Grades PreK K-12	$70 for manual $29 (pack of 25)	Internal Consistency .90 - .97 Test-Retest .80-.91 Interrater: .71-.80	Concurrent Validity BESS to BASC-2 .94 BESS to ASEBA: .76 Construct validity: CFA & EFA find four distinct factors (although only one composite score offered).

Note: Adapted with permissions from Lane, Oakes & Menzies (2010) from the Journal of Disability Policy Studies. SRSS psychometric data were obtained from Lane, Kalberg, Parks, & Carter (2008) and Lane, Parks, Kalberg, & Carter (2007). SDQ psychometric data were obtained from Goodman (1997); Goodman & Scott (1999). SSBD psychometric data were obtained from Walker et al. (1990); Walker, Severson, & Feil, (2010). BERS psychometric data were obtained from Epstein, Harniss, Pearson, & Ryser, (1999); Harniss, Epstein, Ryser, & Pearson (1999). BESS psychometric data were obtained from Dever, Mays, Kamphaus, & Dowdy (2012); Kamphaus & Reynolds (2007); Reynolds & Kamphaus (2004). SSIS-PSG psychometric data were obtained from Gresham & Elliott (2008); Lane et al. (2015).

Table 2

Descriptive Statistics for SSIS and Grades

	Gresham & Elliot (2008)			Current Study				
Measures	N	M	SD	N	M	SD	Skewness	Kurtosis
SSIS-PSG - Prosocial Behaviors	63	3.5	1.1	119	3.68	1.14	-.717	-.190
SSIS-PSG - Reading Skills	63	3.7	1.2	119	3.40	1.31	-.431	-.922
SSIS-PSG - Math Skills	63	3.5	1.1	119	3.36	1.03	-.223	-.379
SSIS Social Skills	63	96.7	14.7	--	--	--		
SSIS Problem Behaviors	63	101.7	13.8	--	--	--		
SSIS Academic Competence	63	98.3	17.4	--	--	--		
Report Card Math	--	--	--	119	80.45	11.48	-1.086	1.225
Report Card Reading	--	--	--	119	81.25	9.40	-.394	-.563
Report Card Conduct	--	--	--	119	3.95	1.15	-1.026	.411

Note. SSIS data were found in the SSIS Rating Scales Manual (Gresham & Elliot, 2008, p. 161) for ages 5-18. SSIS scores are out of a 5 point scale (1 = very limited; 5 = excellent). Report card data were obtained for the current study in grades 1-5. Report card data for math and reading were listed as out of 100 points; whereas, the conduct grade was provided on a 5-point scale (e.g., 1 = unsatisfactory; 5 = excellent).

Table 3

Spearman Rho Correlations (Corrected Where Applicable)

SSIS-PSG Subscales	Report Card Math	Report Card Reading	Report Card Conduct
	r	r	r
Prosocial Behaviors	.424**	.507**	.238**
Reading Skills	.404**	.531**	.229*
Math Skills	.520**	.464*	.119

Note. ** Correlation is significant at the 0.01 level (two –tailed test) * Correlation is significant at the 0.05 level (two –tailed test)
Note. SSIS data were found in the SSIS Rating Scales Manual (Gresham & Elliot, 2008, p. 161) for ages 5-18 (N = 63). Report card data were obtained during the current study for grades 1-5 (n = 161). Report card grades for reading and math are out of 100%; report card grades for conduct were on a 5-point scale with 1 being unsatisfactory and 5 being exceptional.

Table 4

Regression Analyses

Model		Pseudo R^2	Chi-Square		p
Model I – Conduct Grade*					
SSIS-PSG - Prosocial Behaviors		.099	11.575		.021
Model	R^2	B	β	t	p
Model II – Reading Grade*					
SSIS-PSG – Reading Skills	.264	3.627	.514	6.482	.000
Model III – Math Grade*					
SSIS-PSG – Math Skills	.234	5.289	.884	5.983	.000

Note: Model I contains an ordinal dependent variable, therefore ordinal regression was conducted. Models II & III are the result of standard simultaneous regression.

Table 5

Identifying true and false positives for SSIS-PSG ratings and Report Card Grades

	Reading Grades			Math Grades			Conduct Grades		
	< Cut-Off	≥ Cut-Off	Total	< Cut-Off	≥ Cut-Off	Total	< Cut-Off	≥ Cut-Off	Total
Table Key									
Identified by SSIS-PSG	True +	False +		True +	False +		True +	False +	
Not identified by SSIS-PSG	False -	True -		False -	True -		False -	True -	
Total									
SSIS-PSG – Reading Skills									
Identified by SSIS-PSG	10	47	57						
Not identified by SSIS-PSG	3	59	62						
Total	13	106	119						
SSIS-PSG – Math Skills									
Identified by SSIS-PSG				14	53	67			
Not identified by SSIS-PSG				4	48	52			
Total				18	101	119			
SSIS-PSG – Prosocial									
Identified by SSIS-PSG							9	36	45
Not identified by SSIS-PSG							3	71	74
Total							12	107	119

Note: As was used by Kettler et al. (2012), the "cut-off" for the SSIS-PSG is any score of 1, 2, or 3. The "cut-off" for reading and math grades were 69 or below. The "cut-off" conduct grades were "Needs Improvement" or "Unsatisfactory." A True+ is someone who is identified as having a problem (below the cut-off) on both Report Card Grade and SSIS-PSG. A True – is someone who was identified as not having a problem (above the cut-off) when they are meeting the standards on both Report Card Grade and SSIS-PSG.

Table 6

Predicting school grades using teachers' rating of academic and social behaviors

	SSIS-PSG Reading X Reading Grade	SSIS-PSG Math X Math Grade	SSIS-PSG ProSocial X Conduct Grade	Comparison
Sensitivity	10/13=0.76	14/18=0.78	9/12=0.75	Range: .41-.85[2] Target: .70[3]; .90[4]
Specificity	59/106=0.56	48/101=0.48	71/1107=0.66	Range: 81-.97[2] Target: .70[3]; .80[4]
Positive Predictive Value	10/57=0.17	14/67=0.21	9/45=0.20	Range: .25-.58[2]; 26-1.0[5] Target: .75[5]
Negative Predictive Value	59/62=0.95	48/52=0.92	71/74=0.96	Range: 82-1.00[2]; .74-96[5] Target: No Data
False Positive Rate	47/106=0.44	53/101=0.52	36/107=0.34	Range: .20-.60[4]: Target: .90[4]
False Negative Rate	3/13=0.23	4/18=0.22	3/12=0.25	Range: .20-.60[4]: Target: .10-.50[4]
Overall Correct Classification Or Observed Agreement	69/119=0.58	62/119=0.52	80/119=0.67	Range: .62[3]: Target: .90[4]
Chance Agreement	(57)(13)+(62)(106)/ 119^2=0.52	(67)(18)+(52)(101)/ 119^2=0.46	(45)(12)+(74)(107)/ 119^2=0.60	No Data
Kappa	(0.58-0.52)/(1-0.52) =0.13	(0.52-0.46)/(1-0.46) =0.11	(0.67-0.60)/(1-.60) =0.18	No Data
ROC: Area Under the Curve	.743; p = .002	.747; p = .001	.565; p = .477	Standard: .70-.80[4] is fair poor is less than .70[4]

Note[1]: Calculations were conducted based off analyses described in Kessler and Zimmerman (1993).
Note[2]: Based off of a comparison of screeners by Gredler (1997).
Note[3]: Target based on one set for a different screening instrument (Distefano & Kamphaus 2007).
Note[4]: Targets are described in a position paper by Compton and colleagues (2010).
Note[5]: Targets are described in a paper by Milich, Widiger, and Landau (1987)

Figure 1:

ROC Analysis: Reading Skills by Reading Grade

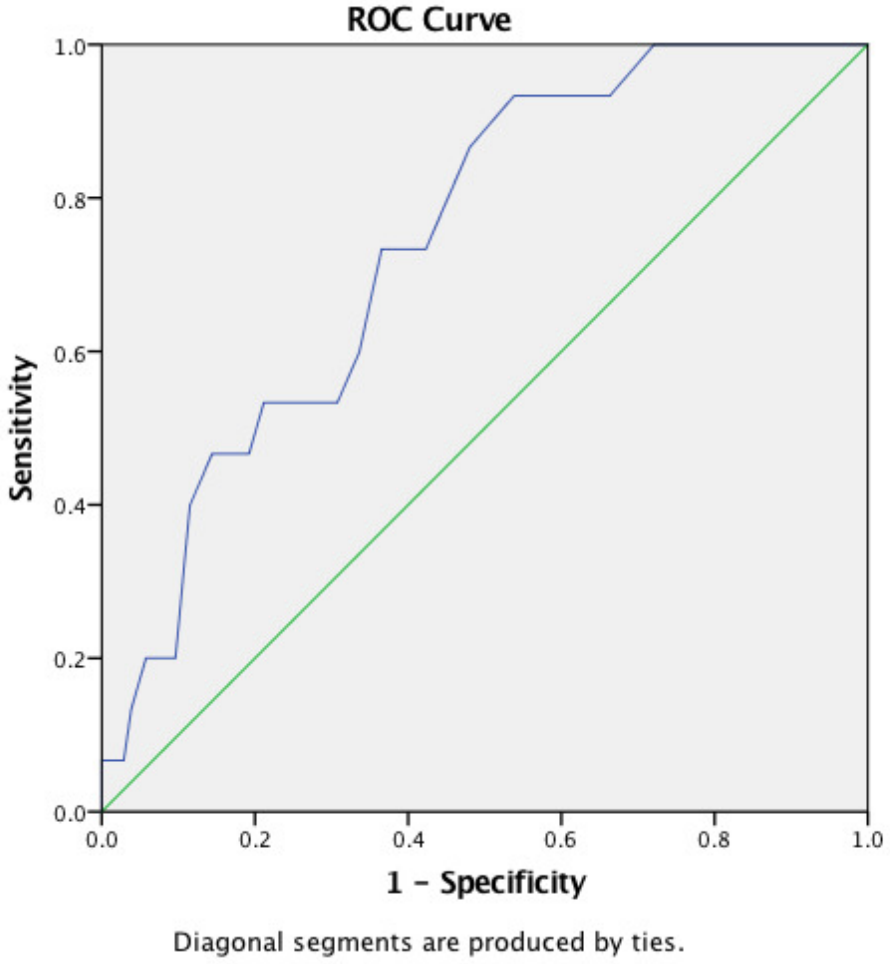

Figure 2:

ROC Analysis: Math Skills by Math Grade

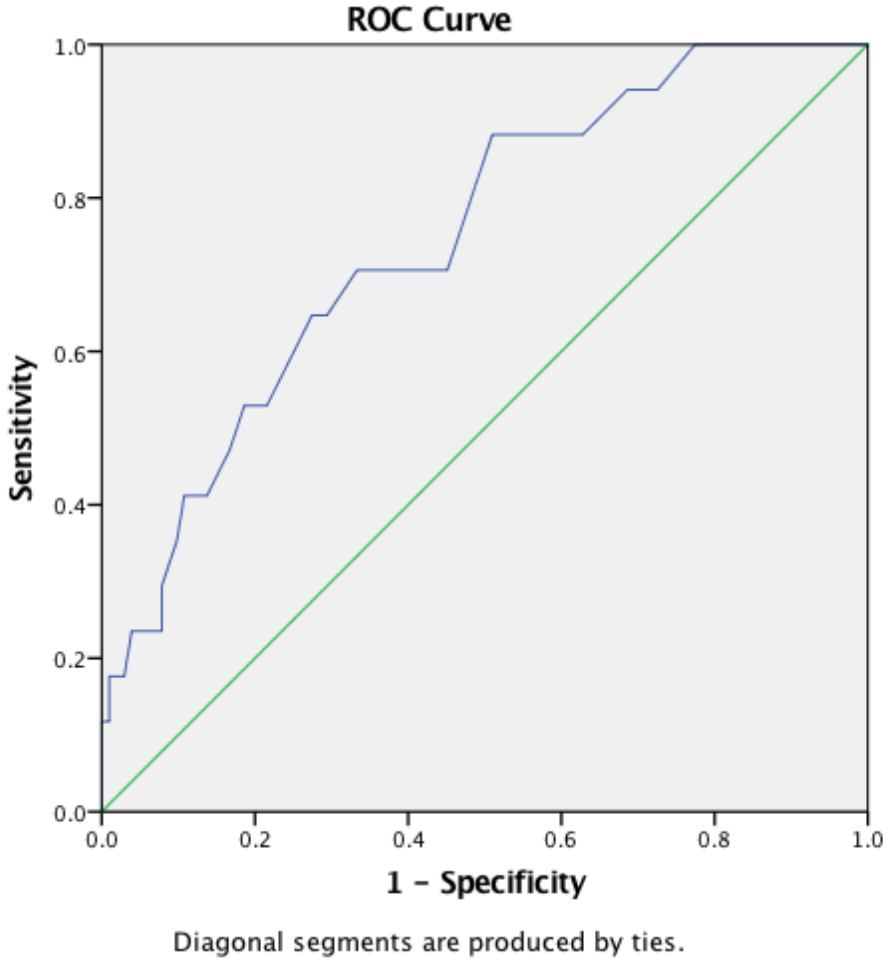

Figure 3:

ROC Analysis: Prosocial Skills by Conduct Grade