UNIVERSAL SCREENING OF STUDENTS' CLASSROOM BEHAVIORS: AN INVESTIGATION OF THE VALIDITY AND CLASSIFICATION ACCURACY OF THE BEHAVIOR SCREENING CHECKLIST III WITH KINDERGARTEN STUDENTS

A Dissertation

by

NICOLE JEAN HALE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Nathan Clemens
Committee Members,	Jamilia Blake
	Lisa Bowman-Perrott
	Lisako McKyer
Head of Department,	Shanna Hagan-Burke

August 2017

Major Subject: School Psychology

ABSTRACT

The purpose of this study was to investigate the Behavior Screening Checklist III (BSC) as a universal behavior screening measure with a sample of kindergarten students. A total of six research questions were examined using bivariate correlation analyses, multiple regression analyses, and receiving operating characteristic (ROC) curve analyses. This was the first study conducted using a ROC curve analyses to investigate classification accuracy indicators of the BSC. Preliminary evidence of the BSC indicates it is a reliable and valid screening measure to use in the schools for identification of students at-risk for later behavior problems. However, the current literature is limited to correlation coefficients and multiple regression coefficients.

Results from the present study concluded the BSC yields consistent teacher ratings over time. Convergent and predictive validity were also supported, using the SDQ as a criterion measure. Multiple regression analyses revealed BSC scores accounted for a statistically significant amount of variance in end of year SDQ scores over and above student ODRs and number of absences. This finding occurred across all three multiple regression models. Winter ORDs were also found to be a statistically significant predictor variable of Spring ODRs. As for classification accuracy, the BSC was able to differentiate between those at-risk from those not at-risk based on AUC. The BSC exhibits good to excellent sensitivity but poor specificity. Of the predictor variables that were statistically significant, Spring BSC scores resulted in the strongest classification accuracy based on the AUC statistics. As this was the first study to

explore classification accuracy of the BSC, it provides emerging support for utilizing the BSC as a universal behavior screening measure.

DEDICATION

I would like to dedicate this dissertation to my mother, who taught me to love life and live it to the fullest. Although she is no longer physically present to celebrate my professional and personal successes with me, her spirit continues to live on and inspire me every day. This is for you, mom.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, advisor, and mentor, Dr. Nathan Clemens. From day one, you never stopped supporting me and pushing me to become a better researcher and school psychologist. Countless conversations and emails that included your words of wisdom and encouragement that helped me get through some of the most stressful times in my doctoral studies. Not to mention, your ability to be candid with me when I needed it the most to stay on track for reaching my goals. Thank you from the bottom of my heart for giving me the opportunity to work with you and learn from you at Texas A&M University. Thank you for always believing in me. And most importantly, thank you for helping me achieve my dream.

I would also like to thank my committee members, Drs. Jamilia Blake, Lisa Bowman-Perrott, and Lisako McKyer for all your support through the process of completing this study. I am forever grateful for your commitment to my success. An additional thank you goes out to the entire staff in the school psychology program at Texas A&M University, including Drs. William Rae, Krystal Cook, and Cynthia Riccio. Each of you have significantly and positively impacted my development as a school psychologist. I can only hope that someday I am able to make an impact on the next generation of school psychologists in-training, the way that you have throughout my doctoral studies. A special shout out goes to Dr. Jeremey Sullivan. Without meeting you, I may have never found my calling to become a school psychologist. Thank you for providing me with the opportunities you did while I began my training at UTSA. I

still remember the first conversation we had in your office regarding the field of school psychology. Little did I know how much that conversation would forever change my life and career.

As for my cohort members, (or as Lindsey's grandmother calls us, "cohoes"), I have no clue where I would be without each and every one of you. We started the program together not knowing each other, but established friendships that are stronger than anything I could have ever imagined. From surviving Dr. Thompson's final exams to late night report writing in Heaton Hall, each one of you have a special place in my heart and will forever be one of Emilia's aunties. All the blood, sweat, and tears that we have shared together over the past four years, definitely makes for one incredible journey at Texas A&M University. All I can say is thank you for your support, encouragement, and enduring friendship.

Last, I would like to acknowledge my family. My in-laws have graciously stood by my side and supported me through my doctoral studies from the moment they walked into my life. Not to mention, the many days my mother-in-law offered to house our dogs for an extended period of time to help reduce my stress levels. Thank you, thank you, and thank you! For my husband, Danny, there are no words that can accurately describe the amount of gratitude I have for you. I simply could not have finished this dissertation without you. I could not have done this without any of your support. All the times I stopped believing in myself, you were right there to catch me and pick me up. Because of your selflessness, I was able to achieve my dream. Someday I hope to do the same for you. I love you from the bottom of my heart. One last acknowledgment goes

out to my daughter, Emilia. You have taught me so much about life and happiness. I never thought I could love someone so much. Although I started my journey well before you came into my life, just know I did all of this for you. I hope to be a role model for you and support you in reaching your goals someday.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor Nathan Clemens [advisor] of the Department of Educational Psychology; Professors Jamilia Blake and Lisa Bowman-Perrott of the Department of Educational Psychology; and Professor Lisako McKyer of the School of Public Health.

The data entry was completed collaboratively by Ashley Smith. The analyses depicted in Chapter 4 were conducted in part by Professor Nathan Clemens of the Department of Educational Psychology. All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a dissertation research fellowship from Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

In 1957, the United States Commission of Chronic Illness proposed to define screening as, "the presumptive identification of unrecognized disease or defect by the application of tests, examinations, or other procedures which can be applied rapidly. Screening tests sort out apparently well persons who probably have a disease from those who probably do not" (Wilson & Junger, 1968, p. 11). This definition was adopted by the World Health Organization in 1968 and is still used in the medical field (Harris, Sawaya, Moyer, & Calonge, 2011). However, one of the first documented forms of screening began in 1917, when the United States Army initiated a program to exclude young men from enlisting with a below average intelligence (Morabia & Zhang, 2004). Since its conception in the military, screening practices have continued to expand across other fields of research such as medicine, community health, psychology, and education.

In education, screening is used to identify students who may be at-risk for experiencing a range of negative consequences in the future (Albers & Kettler, 2014). Implementation of screening practices allow educators to objectively and efficiently monitor the performance of a group of students across a range of skills such as academic achievement or behavior functioning (Dever, Raines, & Barclay, 2012). The main goal of screening is to identify students at-risk and to provide additional intervention supports to prevent the onset of more deleterious symptomology (Chafouleas, Kilgus, & Wallach, 2010). When all students in a school are screened, this practice is referred to as

universal screening and is considered a best practice in the field of school psychology (Albers & Kettler, 2014). In addition, universal screening has the capacity to provide objective insight regarding how well current educational programs are meeting the needs of all students. From identifying students presenting with sub-clinical symptoms to facilitating system-level changes (Albers & Kettler, 2014), universal screening is a valuable practice for school.

Universal screening in the classroom has come a long way since inauguration of the No Child Left Behind Act of 2001 and the Presidential Commission on Excellence in Special Education in 2002. Both federal regulations emphasized the importance of schools engaging in efforts that align with early identification of students experiencing academic and behavioral difficulties, as well as implementation of preventative interventions (Severson, Walker, Hope-Doolitte, Kratochwill, & Gresham, 2007). In response to federal legislation, schools across the nation began to implement problem-solving models to address a variety of academic and behavioral needs (Severson et al., 2007).

In general, a problem-solving model focuses primarily on making data-based decisions for educational programming and planning, as well as implementing evidence-based instruction and support (Bayat, Mindes, & Covitt, 2010). There are three problem-solving models commonly used in schools: Response to Intervention (RTI, primarily academic-focused), Positive Behavior Intervention and Supports (PBIS, primarily behavior-focused), and Multi-Tiered Systems of Support (MTSS, comprehensive). Multi-Tiered Systems of Support is the newest problem-solving model

that integrations key principles of RTI and PBIS (McIntosh & Goodman, 2016). Some refer to MTSS as a comprehensive problem-solving model, as it targets academic, social, emotional, and behavioral needs of all students (McIntosh & Goodman, 2016). Regardless of the type of problem-solving model used, they generally follow a three-tiered service delivery for identifying students' in need of individualized interventions and defining the level of warranted intervention intensity (Bayat et al., 2010).

Tier 1 is often referred to as the primary or universal tier, which provides all students with evidenced-based instruction. In Tier 1, schools are engaged in continuous monitoring of progress by screening all students (Bradley, Danielson, & Doolittle, 2007). It is known that approximately 80% of students will respond appropriately to universal instruction and support. Students who were previously identified at-risk during Tier 1 universal screening, typically are not responding to universal intervention efforts and may benefit from more intensive supports. These students move up the hierarchy and are placed in Tier 2. Tier 2, involves providing a smaller group of students with additional support or interventions (Bayat et al., 2010). It is expected that 10 to 15% of students will fail to respond to universal supports and need more intensive intervention. Tier 3 is considered the most intensive level of instruction and support. It is estimated that approximately 5 to 10% of students will not respond to Tier 1 or Tier 2 efforts, and need Tier 3 support. Across all three tiers, it is best practice for schools to engage in continuous data collection and progress monitoring to determine how each student is responding (or not responding) to prevention and intervention efforts (Bayat et al., 2010). According to Fuchs and Fuchs (2007), students who fail to respond to all

three tiers should then be referred for a comprehensive evaluation to determine if an educational disability is present.

Research has documented several benefits for schools implementing a multi-tiered problem solving model with fidelity (McIntosh & Goodman, 2016). Several studies have documented a reduction in bullying and disruptive behaviors in the classroom, as well as an increase in academic achievement, emotional and self-regulation in students, an improved school climate and positive student-teacher relationships (Bradshaw, Koth, Thorton, & Leaf, 2009; Kelm & McIntosh, 2012; Waasdorp, Bradshaw, & Leaf, 2012). When students who are experiencing academic and behavioral challenges are identified early, the likelihood increases that they will receive appropriate treatment to remediate maladaptive trajectories and offset the course of negative outcomes (Gresham, Hunter, Corwin, & Fischer, 2013).

Prior to multi-tiered service delivery models, student failure was the primary signal to initiate intervention, which significantly delayed intervention or treatment (Gresham, 2008). The longer intervention services are delayed, the more resistant maladaptive behaviors are to positive change. When students are identified early in their development, behaviors are often mild and more apt to respond to intervention, as compared to waiting later in life when the behaviors may be more chronic or severe (Gresham et al., 2013). Universal screening practices have also been shown to reduce referral bias, as progress is documented through various objective measures that are observable (Hawken, Vincent, & Shumann, 2008). Finally, Gresham et al. (2013) discussed how tiered models of service delivery have shifted the focus away from the

individual student as being the problem and encourages change within the environment.

Universal Screening for Behavior

For the past several years, researchers have laid the groundwork to support the effectiveness, value, and process of screening for academic achievement, as many schools are now engaged in some type of academic screening (Dowdy et al., 2014). Unfortunately, early detection of problem behavior in the classroom has not received the same level of attention as academic screening, resulting in a large gap in the research (Dowdy et al., 2014). Severson and colleagues (2007) discussed how some of the resistance to pursue universal behavior screening is rooted in the historical ideology that schools' sole responsibility is to foster academic growth and not behavior growth. In the past, many teachers believed that behavior problems were remediated through initiation of the refer-test-place process (e.g. special education evaluation and eligibility), providing students with services only after the behavior problems have become chronic and resistant to change (Severson et al., 2007). Furthermore, it is estimated that only 2% of schools engage in some type of emotional and behavioral screening in the classroom (Romer & McIntosh, 2005).

Without universal screening practices, early identification and intervention services are often delayed, allowing the opportunity for behavior problems to persist long enough to reach the threshold of a clinical disorder (Lane, Jolivette, Conroy, Nelson, & Benner, 2011). Students presenting with early symptoms of behavior problems are at a much greater risk for experiencing higher rates of school suspension; more likely to be placed in a restrictive setting or an alternative learning environment;

are at greater risk of school dropout (approximately 53% in 2013); accumulate lower wages as an adult; are more likely to be unemployed; experience higher rates of reliance on welfare benefits; and damaged relationships with family, significant others, and peers (Gresham et al., 2013). Additionally, there is a higher probability that individuals with early behavior problems will be incarcerated, engage in risky behaviors such as sexual activity and/or drug use, and display physical violence (Eber & Nelson, 1997; Ryan, Reid, & Epstein, 2004). Given the substantial evidence regarding poor prognosis associated with early onset of emotional and behavioral disorders, universal screening is critical for early identification of students who may be at-risk in order to provide early interventions (Dowdy et al., 2014).

Universal Behavior Screening Measures

For schools to engage in universal screening practices and to accurately identify students who may be at-risk, several screening measures have been developed. These include, but are not limited to, the Behavior and Emotional Screening System (BESS; Kamphaus & Reynolds, 2007); the Systematic Screening for Behavior Disorders (SSBD; Walker & Severson, 1992); the Student Risk Screening Scale (SSRS; Drummond, 1994); the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997); and the Behavior Screening Checklist III (BSC; Muyskens, Martson, & Reschly, 2007). In addition, schools frequently use other sources of data readily available to them, such as office discipline referrals (ODRs), suspensions, academic achievement, performance on state standardized testing, and student attendance records. Across the different universal screening measures, there are several studies published that established the reliability

and validity of each measure (see Chapter II for references regarding each screening measure). However, in reviewing the available research on universal screening measures, very few have investigated classification accuracy.

Classification accuracy is defined as an assessment or screening measures' ability to differentiate between individuals who may not be at risk (true negatives) from those who are at risk (true positives; Glover & Albers, 2007). Classification accuracy is characterized by values of sensitivity, specificity, positive predictive values (PPV), and negative predictive values (NPV; Glover & Albers, 2007). Inadequate classification accuracy of a screening measure is potentially dangerous, as it may incorrectly identify students as "at-risk", when in fact they are not or vice-a-versa. Misidentification and inaccurate labeling may lead to a range of negative consequences such as loss of resources, time, and money used for supporting those at-risk (Glovers & Albers, 2007). More concerning are situations in which screening tools fail to identify students who are truly at-risk, thus leading to missed opportunities for providing early intervention and increasing the possibility of developing chronic behavior patterns that become resistant to change.

Studies that have investigated classification accuracy of universal screening measures for emotional and behavioral difficulties are limited in number and scope (Cook et al., 2011). As with any new screening measure it is imperative to establish the reliability and validity; however, this should also include classification accuracy (Glover & Albers, 2007). Glover and Albers (2007) extensively reviewed considerations for selecting universal screening measures, of which they emphasized classification

accuracy. Since the release of the Glover and Albers (2007) article, it appears that more studies have included classification accuracy in their analyses. A systematic literature search was conducted and found no studies prior to 2007 that investigated classification accuracy using universal behavior screening measures in the context of a school setting.

Over the past ten years, research has been published focusing on exploring the classification accuracy of universal screening measures. The Student Risk Screening Scale (SRSS; Drummond, 1994) has been used in several studies, using a receiving operating characteristic curve analysis to investigate the sensitivity and specificity (Ennis, Lane, Oakes, 2012; Lane et al., 2009; Lane, Kalberg, Lambert, Crnobori, & Bruhn, 2010). All three studies yielded comparable results, as the SRSS demonstrated stronger classification accuracy for externalizing behaviors compared to internalizing behaviors. These results also determined the SRSS has stronger classification accuracy when compared to the Student Screening for Behavior Disorders (Walker & Severson, 1992). The Behavioral and Emotional Screening System (BESS, Kamphaus & Reynolds, 2007) is another well, established measure that have reported classification accuracy statistics (DiStefano & Kamphaus, 2007). Feil, Severson, and Walker (1995) explored the development of the BESS and conducted a receiving operating characteristic curve analysis using the Early Screening Project (ESP, et al.,1995) as the criterion measure. Based on the cut-scores selected, an area under the curve was 0.70 (e.g. 70%) probability of correctly discriminating between students at-risk from students who are not at-risk (DiStefano & Kamphaus, 2007). As these are only a few studies, most universal screening measures have not established classification accuracy.

The Behavior Screening Checklist III is one example of a universal screening measure that is relatively new and has limited research published (Muyskens et al., 2007). This screening measure was initially developed out of the Minnesota Public school system as a guide to behavior consultation in the classrooms (Muyskens et al., 2007). There are several advantages to using the BSC for universal screening. First, the screening measure only includes 12 items and takes less than a minute to complete per student, compared to other screening measures which can take upwards of 5 to 10 minutes per student. In addition to time, the cost to use is free. This is even more appealing to schools, as they are often limited in their financial spending and available resources. The BSC also screens for a range of classroom behaviors that are often indicative of later academic failure or behavior difficulties. To date, only two studies have been published on the BSC, both of which have explored reliability and validity. Preliminary data provides evidence to support the psychometric properties of the BSC, yet no studies have investigated the classification accuracy. Due to the limited number of studies conducted using the BSC, there are several possible research questions to explore.

Study Purpose

The purpose of this study was to investigate the psychometric properties of the BSC when used as a universal behavior screening measure with a sample of kindergarten students. Convergent and predictive validity was explored using the SDQ as the criterion measure. Multiple regression was also used to identify what variables accounted for the greatest amount of variance in year-end scores on the SDQ. In

addition, this study investigated the classification accuracy of the BSC when it comes to identify students at-risk for experiencing a range of challenging classroom behaviors. A total of six research questions were explored:

1. How consistent are the BSC Total scores across administration with a sample of kindergarten students? It was hypothesized that BSC Total scores would yield strong and positive correlation coefficients across administration. Investigation of correlation coefficients focused on Fall BSC Total scores and Winter BSC Total scores; Winter Total BSC scores and Spring BSC Total Scores; Fall BSC Total scores and Spring BSC Total scores.
2. To what degree does each administration of the BSC demonstrate convergent validity with the SDQ? It was hypothesized that Fall BSC Total would yield a strong and positive correlation coefficient with Fall SDQ Total Difficulties. It was also hypothesized that similar results would be obtained between Winter BSC Total and Winter SDQ Total Difficulties, as well as the Spring BSC Total and Spring SDQ Total Difficulties. It was also hypothesized that BSC Classroom Behavior would strongly and positively correlate with all subscales on the SDQ (except the Prosocial Behavior subscale, this should be negatively correlated). The Externalizing Behaviors subscale on the BSC would strongly and positively correlate with the Hyperactivity/Inattention and Conduct Problems subscales on the SDQ. As for the BSC Socialization subscale, it was hypothesized these scores would strongly and negatively correlate with the Prosocial Behavior subscale on the SDQ.
3. To what degree do the Fall BSC and Winter BSC scores demonstrate predictive

validity with the SDQ? It was hypothesized that Fall BSC Total would strongly and positively correlate with Spring SDQ Total Difficulties. It was also hypothesized that Winter BSC Total would strongly and positively correlate with Spring SDQ Total Difficulties.

4. What proportion of variance in Spring SDQ Total Difficulties are accounted for by BSC Total, student absences, and ODRs? This research question included the exploration of three different models to capture data collected at each time point (fall, winter, and spring model). It was hypothesized that BSC Total scores (at each time point) would account for a statistically significant amount of variance in Spring SDQ Total Difficulties. It was also hypothesized that ODRs and student absences would not count for a statistically significant amount of variance in year-end SDQ scores.
5. Does BSC Total scores demonstrate acceptable classification accuracy in identifying students at-risk for behavioral challenges in the classroom? It was hypothesized the BSC Total scores across all three administrations would demonstrate acceptable classification accuracy.
6. Does the BSC Total scores demonstrate stronger classification accuracy in identifying students at-risk when compared to ODRs and student absences? Only variables identified as a statistically significant predictors in the multiple regression models will be used to answer this research question. It is hypothesized that the BSC will demonstrate stronger classification accuracy, over and above ODRs and student absences.

CHAPTER II

REVIEW OF THE LITERATURE

Classrooms are filled with students who come from all walks of life. As students enter their classrooms, they present with a wide range of academic skills, along with even more diverse ethnic, social, religious, cultural, and socioeconomic backgrounds than we have ever seen before in the public education system. As schools continue to embrace this level of diversity in classrooms, teachers are placed in a unique, yet challenging position that requires greater differentiation of academic instruction, while at the same time meeting the social, emotional, and behavioral demands of individual students (Lane, Menzies, Oakes, & Kalberg, 2012). To put in perspective, approximately 20% (or 1 in 5) of children and adolescents will, at some point, experience significant emotional and behavioral impairment to the degree that warrants a clinical diagnosis (Jaffee, Harrington, Cohen, & Moffit, 2005; U. S. Department of Health and Human Services, 1999). Despite the high prevalence rates of emotional and behavioral difficulties in our school-aged children, less than 1% are identified as a student with an Emotional Disturbance (ED) under the Individuals with Disabilities Education Improvement Act of 2004 (IDEA; Lane et al., 2010). The number of students identified under the disability category of ED has also remained relatively stable since 1986, which places a question mark next to the current practices that are in place for identifying students with emotional and behavioral difficulties (Nordness, Epstein, Cullinan, & Pierce, 2014).

When children and adolescents go unidentified and untreated for emotional and behavioral difficulties, research has shown this often presents as a barrier to learning, resulting in academic failure and a wide range of problems later in life (Catalano, Haggerty, Osterle, Fleming, & Hawkins, 2004). A longitudinal study conducted by Bradley and colleagues (2008), found of students who experienced emotional and behavioral difficulties, 75% achieved below grade levels in reading and 97% achieved below grade levels in math. In addition, individuals presenting with early symptoms of emotional and behavior problems are also at risk for experiencing higher rates of school suspension; they are more likely to be placed in an alternative learning environment; are at greater risk of school dropout (approximately 53% in 2013); accumulate lower wages as an adult; are more likely to be unemployed; experience higher rates of reliance on welfare benefits; and have damaged relationships with family, significant others, and peers (Gresham et al., 2013; Reinke, Herman, Petras, & Ialongo, 2008). Furthermore, there is a higher probability that individuals with early onset of emotional and behavioral problems will be incarcerated, engage in risky behaviors such as sexual activity and/or drug use, and display physical violence (Eber & Nelson, 1997; Ryan et al., 2004). Despite the large percentage of students who experience emotional and behavioral difficulties, a significant number of these students are left unidentified (Eklund et al., 2009). As a result, these students do not receive the necessary interventions to remediate their symptoms (Eklund et al., 2009).

In the past, schools have attempted to remediate the number of students experiencing emotional and behavioral difficulties in the classroom through a model

commonly known as “refer-test-place.” With this model, teachers are largely accountable for identifying students in need of additional support services and referring them for a comprehensive special education evaluation (Kamphaus, DiStefano, Dowdy, Eklund, & Dunn, 2010). The problem with this identification process is students who are presenting with the greatest level of impairment in the classroom are often the ones who would be referred for an evaluation and receive services (Kamphaus et al., 2010). With this model, students are not effectively or efficiently identified for additional support services (Tilly, 2008). This approach is also considered reactive, meaning that students may not be identified for additional supports in the classroom until after emotional and behavioral functioning have led to significant impairment. Tilly (2008) also acknowledged how teachers are equipped with a range of teaching abilities when it comes to working with struggling students, which directly impacts the referral rate. In addition, many educators do not receive formal training on how to identify emotional and behavioral difficulties with their students (Tilly, 2008). To address limitations of the refer-test-place model, multitiered models of instruction and support (e.g. Multi-Tiered Systems of Support) have been developed and implemented in schools (Martson, 2002).

Multi-Tiered Systems of Support (MTSS) is a problem-solving model that utilizes continuous assessment data to identify how well students are responding to universal instruction and support (Muyskens et al., 2007). When implemented with fidelity, MTSS is designed to support all students with various academic skills, as well as emotional and behavioral needs. Multi-Tiered Systems of Support is grounded in data collection, progress monitoring, and continuous data analysis. The overarching goal of

MTSS is for schools to effectively and efficiently identify students in need of more intensive instruction and support. When it comes to students experiencing emotional and/or behavioral difficulties, there are often early signs or symptoms indicative of future impairment (Cullinan & Epstein, 2013).

Early identification increases the likelihood of receiving early intervention services (Eklund et al., 2009). Early intervention has been shown to reduce or minimize symptom severity and intensity, offset the developmental trajectory for maladaptive behaviors, and prevent other symptoms from developing (Eklund et al., 2009). In order for schools to accurately identify students who necessitate additional support and intervention services, this requires schools have access to quality screening measures that are "contextually appropriate, technically sound, and usable" (Glover & Albers, 2007, p. 118). Glover and Albers (2007) wrote their article with the intention of providing researchers and practitioners with some guidance on quality indicators of universal screening measures and what factors should be considered prior to administration. These characteristics may be applied to any type of assessment but are reviewed below within the context of universal screening.

Quality Indicators of Screening Measures

Evaluation Considerations

The first characteristic that Glover and Albers (2007) discussed involves determining whether or not the universal screening tool is appropriate for use within the context of administration. There are essentially two types of universal screening measures available for use in the schools: 1) those that predict the students who are at-

risk for developing difficulties in the future or 2) those that identify students who are currently experiencing difficulties in the classroom. When selecting a universal screening measures, this should be one of the first considerations. Are the screening measures being utilized to identify students in need of preventative services (Tier 2) or is the goal to capture students who are in need of more intensive intervention services (Tier 3)?

Glover and Albers (2007) suggest many of the universal screening measures fall along this continuum; however, selection of a measure that identifies students across various levels of risk may be more efficient and effective at serving all students. In addition, universal screening measures should be grounded with theoretical and/or empirical support that align with constructs of interest. Schools should select universal screening measures that are contextually and developmentally appropriate for their student population. For example, a screening measure that was developed to use with an adult psychiatric population to screen for psychosis would not be appropriate for use in a school.

Technical Adequacy

Glover and Albers (2007) provided an in-depth review of the psychometric properties that are critical for a universal screening measure to be considered technically sound or adequate. They referenced *The Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Center on Measurement in Education [NCME], 1999), a comprehensive review of guidelines for determining technical adequacy of an

assessment tool. Glover and Albers (2007) specifically highlighted the importance of the standardization (e.g. normative sample), measurement consistency (e.g. reliability), and accuracy in identification of those at-risk (e.g. validity).

Standardization

The standardization of a universal screening measure involves consideration of the normative sample, whether this was done locally or nationally. National norms are used to offer information regarding the level of impairment or functioning compared to same age or grade level expectations. National norms also provide some added benefit as they are typically more stable and less dependent on environmental factors. However, local norms offer information that may not be available when using national norms. Local norms offer meaningful implications for intervention as they tend to capture a robust representation of the individuals screened. Whether national or local normative samples are used in the standardization process "representativeness, recency, and sample size" are vital when determining the adequacy of the norms (Glover & Albers, 2007, p. 122). Aside from the norming sample used, there are several additional statistics that should strongly be considered and reviewed prior to using a universal screening measure. Those include reliability, validity, criterion-related validity, concurrent validity, construct validity, content validity, and predictive validity.

Measurement Consistency

When looking to establish measurement consistency, reliability is typically evaluated. Reliability is a term used to describe how similar or consistent the performance of a universal screening measure (or any assessment tool) is over time and

between different raters (Salvia & Ysseldyke, 2004). Reliability coefficients range from 0.00 to 1.00 and larger coefficients are indicative of better reliability. Reliability involves internal consistency, test-retest reliability, and inter-rater reliability (Glover & Albers, 2007). Internal consistency captures how well items on an assessment or a subscale measure the same construct. Test-retest demonstrates the performance of a measure over time, when administered to the same individual. Last, inter-rater reliability refers to the relationship of the universal screening results between two different raters. All three types of reliability should be evaluated in order to understand how consistent the universal screening measure performs over time and between different raters, prior to data collection.

Accuracy of Risk Status

Validity refers to the accuracy of an assessment tool or a universal screening measure. Glovers and Albers (2007) refer to *The Standards for Educational and Psychological Testing* (AERA et al., 1999) for a detailed review of the various types of validity but explore the following categories of validity for the purpose of universal screening measures: construct validity, content validity, and criterion-related validity. Construct validity refers to the degree to which the assessment actually measures the constructs purported to measure. Content validity, which is often confused with construct validity, refers to the actual items on the screening measure as they relate to the underlying constructs. For the purpose of this study, an in depth exploration of criterion-related validity is discussed.

Criterion-related validity is a characteristic important for establishing technical adequacy of a universal screening measure because it is "an indicator of how well an assessment predicts an individual's performance on a specified criterion" (Glover & Albers, 2007, p. 123). This is typically established through evaluating the correlation coefficients between the universal screening measure and a gold-standard measure. Methods of establishing criterion-related validity include concurrent validity and predictive validity. When it comes to universal screening measures, concurrent validity refers to the relationship between scores on an established measure (criterion) and a universal screening measure when both measures are administered at the same time.

Although screening tools should be consistent over time (reliability) and correlated with important outcomes of interest (validity), universal screening measures must also demonstrate adequate classification accuracy and be able to distinguish between individuals who will not have difficulties from those who will. Inadequate classification accuracy of a screening measure is potentially dangerous, as it may incorrectly identify students as "at-risk", when in fact they are not (or vice-versa). Misidentification and inaccurate labeling of students may lead to several negative consequences including a loss of resources, time, and money used for supporting students not in need of additional support (Glovers & Albers, 2007). More concerning are situations in which screening tools fail to identify students who are actually at-risk, thus leading to missed opportunities for providing supplementary support and increasing the possibility of developing chronic behavior patterns that become resistant to change. Classification accuracy is established through four separate categories known as

sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Sensitivity is used to describe how accurately a screening measure is at detecting a disorder when it is actually present (Harber, 1981). Sensitivity is calculated by dividing the total number of true-positives from the total number of positive cases that actually exist (Pintea & Moldovan, 2009). Positive predictive power is related to sensitivity, and represents the probability that an individual who is identified as "at-risk" on the screener truly has the disease, disorder, or condition of interest (VanDerHeyden, 2011). Alternatively, specificity is used to describe how accurately a screening measure is at detecting when a disorder is not actually present (Harber, 1981). Specificity is calculated by dividing the number of true-negative cases observed by the total number of negatives that actually exist (Pintea & Moldovan, 2009). Negative predictive power is related to specificity, and represents the probability that an individual who is identified as "not at-risk" on the screener truly does not have the disease, disorder, or condition of interest (VanDerHeyden, 2011). Sometimes screening measures will be wrong and misclassify individuals, which is referred to false-positive and false-negative rates (Fawcett, 2006). A false-positive is an individual identified as at-risk by the screener, when in fact they are not. On the other hand, a false-negative is an individual not classified at-risk by the screener, when in fact they are at-risk (Fawcett, 2006). To summarize these concepts regarding classification accuracy, Table 1 is commonly used to facilitate interpretation and calculations of decision-making (Pintea & Moldovan,

2009). Table 1 provides four different outcomes that explain the classification accuracy of a screening measure.

Table 1

Summary of Classification Outcomes

		Actual Outcome	
		Positive *(disorder present)*	Negative *(disorder not present)*
Screening Results	Screening Test (+) *"At-risk" result*	True Positives	False Positives
	Screening Test (-) *Not "at-risk" result*	False Negatives	True Negatives

Overall, Glover and Albers (2007) argue that although all four indices (sensitivity, specificity, PPV, and NPV are important to determine the technical adequacy of a universal screening measure, sensitivity and positive predictive values should be given priority. Universal screening tools with high sensitivity are preferred over high specificity because it is ideal to capture those who are at-risk or those who are currently experiencing emotional and/or behavioral difficulties in the classroom. They further argue with the consistent under-identification of students who need additional emotional and behavioral support in the classroom, a screening measure that has lower PPV may help address this issue. However, this also needs to be balanced with available resources the school has to offer for support and intervention services once students are identified. It is recommended that during the first phase of universal screening, the

measure demonstrates high sensitivity probabilities. Any measures used to follow up should have high PPV.

Usability

The last characteristic Glover and Albers (2007) discussed in their article refers to usability of a universal screener. They argue that a universal screening measure may be contextually appropriate and demonstrate technical adequacy, but it also must be reasonable to administer within the context of a school setting. Usability includes at minimum the following considerations: cost/benefit analysis; feasibility of implementation; buy-in from multiple stakeholders such as school personnel, parents, students; infrastructure for collecting, managing, and interpreting the screening data; availability of appropriate accommodations with the screening measure; and screening data that is useful and improves treatment decisions. With strict budgets that schools are often faced with, a cost/benefits analysis is often a top priority. The cost of administering the universal screening measure should not outweigh the benefits. Examples of costs includes things such as taking away from instructional time or significant burden on financial resources. As previously mentioned, a universal screening measure that has a lower positive predictive value will over-identify students. If a school has limited resources for providing intervention and treatment, this would be a cost/benefit analysis that should be strongly considered.

Timing and Frequency

As research continues to expand on understanding characteristics and considerations of universal behavior screening, one area that requires additional attention is the timing and

frequency of screening assessment. Dowdy and colleagues (2014) reported the timing and frequency of universal screening practices is one of the "major uncertainties" (p. 454). Some studies support the idea that screening should occur multiple times over the course of a year and include every student in the school (Chafouleas, et al., 2010; Ennis, Lane, & Oakes, 2012). On the other hand, some recommend collecting screening data only during major developmental periods that align with "institutional time points" (Dowdy et al., 2014, p. 454), which would include transitory periods in students' life such as entering kindergarten or moving from elementary school to middle school. This approach assumes specific time points in a student's developmental experiences may cause higher levels of emotional distress and present early risk factors that could be targeted with early intervention. In addition, it has been recommended to collect screening data to assess all students at the beginning of the school year (Stoep et al., 2005) or only when concerns arise (Dowdy, Furlong, Eklund, Saeki, & Ritchey, 2010). However, one could argue that waiting until concerns arise would not be considered screening.

Researchers have also argued that emotional and behavioral symptoms are often transitory during childhood or early adolescences; therefore, frequent screening and monitoring is recommended (Dowdy et al., 2014). However, Dowdy and colleagues (2014) conducted a longitudinal study and their findings suggest otherwise. The focus of their study was to determine if participants emotional and behavioral risk status changed within one year intervals over the course of four years. Their results determined that students risk status was relatively stable over time. They found

moderate to large stability coefficients, with larger stability coefficients between shorter intervals of time. Interestingly, their findings were also stable across risk status for internalizing and externalizing disorders, which is contrary to the current literature that supports instability of risk status for internalizing disorders (Dowdy et al., 2014). The only change that was noted in risk status occurred with the students who were initially identified as having the greatest risk, as they transitioned from "elevated" to "extremely elevated" (p. 465). Results from this study suggest that risk-status is not as transient as researchers once thought (particularly with internalizing disorders).

Dever, Dowdy, Raines, and Carnazzo (2015) found similar results in their study that evaluated risk status over a two-year period. Results concluded that students remained in a similar risk status and variables such as gender, race/ethnicity, socioeconomic status, grade level, transition between schools, and special education status were not predictive of movement. These findings were consistent across internalizing and externalizing problems. Dever et al. (2015) also found that students who fell within the average or normal range during the initial screening, continued to remain within the normal range two years later. However, some transition did occur within the at-risk sample. For example, they found that 40% of students who were identified at-risk during the first screening remained at-risk during the second screening. The researchers acknowledge they were unaware if any interventions had been delivered to account for 60% of participants to change from at-risk to normal over the course of two years. Both of these studies shed some additional light surrounding the questions related to frequency and timing of administering universal screeners in a school.

However, these results are preliminary and future research needs to continue exploring this area to develop a better understanding of timing and frequency.

Available Emotional and Behavioral Screening Measures

To date, there are several emotional and behavioral rating scales that are commonly used for psychological assessment, such as the Child Behavior Checklist (Achenbach & Rescorla, 2001), the Social Skills Improvement System (Gresham & Elliot, 2008), the Behavior Assessment System for Children, Third Edition (Reynolds & Kamphaus, 2015), and the Behavioral and Emotional Rating Scale (Epstein, 2004). These measures are not appropriate for universal screening practices if following Glover and Albers' (2007) recommendations. Length of time to complete each of these measures and cost to purchase per student make them impractical for large-scale, universal screening. In fact, Dever et al. (2012) argued that many school-based practitioners frequently experience uncertainty when it comes to selecting universal behavior screening measures that appropriately meet their schools' needs and goals. This may be the result of not having access to current research on available screening measures. Screening measures that are appropriate for universal practices have been grouped into three separate categories: multiple-gating screening, universal screening, and data that inform screening practices. Examples of each category are reviewed, along with highlights of the current research.

Multiple-Gate Screening Measures

Walker, Small, Severson, Seeley, & Feil (2014), defined multiple-gating as "a generic process involving multiple assessments that cost efficiently identify a subset of

individuals from a larger pool of target participants with a combination of methods and measures generally arranged in sequential order" (p. 47). Multiple-gate screening practices were originally developed by Cronbach and Gleser in 1965 and since have been identified as a best practice in screening assessment (Walker et al., 2014). One of the documented benefits to using this approach is how it is considered cost-efficient and effective at screening a large number of individuals (Dowdy, Dever, Raines, & Moffa, 2016). In the first gate, all students have an equal chance of being identified for passing through to the next gate of assessment (Dowdy et al., 2016). However, this can be argued against because teachers are rank order their students and do actually complete a behavior screening measure on each child in the classroom.

Multiple-gate is a screening process that is extremely dependent on teacher judgment and relies heavily on knowledge of behaviors being screened for (Walker et al., 2014). This type of screening runs a high risk of not identifying students who are experiencing internalizing disorders or identifying only those who present with the most significant externalizing problems. Research suggests teachers have a more difficult time identifying students at risk for internalizing disorders compared to externalizing disorders (Dowdy et al., 2016). Results of the first stage may be highly influenced by teacher bias, which impacts the identification of students who receive additional screening in the second gate (Walker et al., 2014). Other concerns related to multiple-gate screening involves identifying how many gates are ideal, selecting informants to use, as well as acceptable rates of sensitivity and specificity levels at each gate (Dowdy et al., 2016). A few examples of multiple-gate screening measures are described below.

Systematic Screening for Behavior Disorders (SSBD)

The SSBD was developed by Walker and Severson (1992) and is a commonly known multiple-gate screening measure used to monitor prosocial behavior for students in first through ninth grade. The multi-gate format includes three levels (or tiers) in which teachers first nominate six students who they believe may be at-risk. They are asked to identify three students who present with the most concern regarding externalizing behaviors and three with internalizing behaviors. Next, teachers rate adaptive and maladaptive behaviors for only those students nominated. Last, a structured observation is conducted as a follow up for those students who receive behavior rating scores deemed at-risk. The SSBD is available online to purchase from Northwest Publishing Company and the total cost is $550 per school for a 12 month subscription. There is an option to purchase the protocols and manual as a hardcopy or digital. Additional forms cost $10 per classroom.

Walker and Severson (1992) reported acceptable estimates of internal consistency, test-retest reliability, and concurrent validity. Furthermore, the SSBD has been established as an appropriate screening measure to use with elementary students and is able to accurately differentiate between students presenting with internalizing and externalizing symptoms from students who are considered not at risk (Walker & Severson, 1992). Despite the promising psychometric properties of this measure, the SSBD has been widely criticized by the length of time to complete all three levels of the gating system (Dever et al., 2012). In addition, the number of students that can pass

through first gate may limit teachers who have more than six students with at-risk behaviors (Lane et al., 2012).

The methodology supporting the SSBD suggests that prevalence of behavioral problems is the exact same in all grades, schools, and classrooms (Dever et al., 2012). Lane et al. (2012) discussed how the SSBD does not take into account for comorbidity. Students experiencing externalizing disorders are also more likely to experience comorbid internalizing disorders. When teachers recommend that students pass through the first gate, they can only place students on either externalizing or internalizing behavioral dimension. This factor alone may limit the type of services students have access to or potentially impede progress if they are receiving treatment for externalizing problems when internalizing symptoms should also be targeted.

Early Screening Project: A Proven Child Find Process (ESP)

The ESP is another multiple-gate screening measure used with children between the ages of 3 and 5. This measure was developed and modeled after the SSBD. The ESP "incorporates the proactive, universal screening standards for emotional and behavioral problems required by Head Start Performance Standards and ensures that all children are screened for the presence of these problems" (Feil, Walker, Severson, & Ball, 2000, p. 14). This screening measure involves three different gates, each with progressively more intensive assessment procedures with the goal of identifying those students who require early prevention interventions that target emotional and behavioral symptomology (Lane et al., 2012). The manual and measures (both in English and

Spanish) are available online for free at https://research.ori.org/esp/resources.php, once contact information is provided.

Operational definitions, examples, and nonexamples are given for both externalizing and internalizing behaviors (Lane et al., 2012). Teachers then nominate the top five students whose behaviors most match the criteria given for externalizing behaviors and the top five students for internalizing behaviors. Once this process is completed, teachers then rank order the five students from "most likely" to "least likely" for each category of behavior. The students ranked in the top three for externalizing and the top three for internalizing pass through to the second gate. During the second gate, the teacher completes four different behavior ratings scales for each student who were nominated through the first gate. Each of the measures used during the second gate are nationally normed with cut-scores to determine risk status. The third gate involves direct observations and parent questionnaire of the students who passed through gates one and two (Lane et al., 2012).

Research has provided substantial evidence supporting the ESP as a reliable and valid screening measure that is also practical to use within the early elementary school setting (Feil, Severson, & Walker, 1998). Interrater reliability, test-retest stability, as well as content, concurrent, and discriminative validity have all been established (Feil, et al., 1995). Previous studies have explored and confirmed adequate convergent validity with the Preschool Behavior Questionnaire (Behar & Stringfield, 1974); the Conners' Teacher Rating Scale (Conners, 1989); the Achenbach's Child Behavior Checklist – Teacher Report Form (Achenbach & Edelbrock, 1987); and Social Skills Rating System

(Gresham & Elliot, 1990). Overall, there are several studies that have explored the ESP to establish and document the psychometric properties.

Universal Screening Measures

Universal screening measures used in the school environment involve teachers or staff completing a rating scale for every student in their respective classroom (Eklund et al., 2009). Some screening measures also include a parent rating scale or a student self-report (e.g. Strength and Difficulties Questionnaire). Universal screening measures are developed to assess a broad range of emotional and behavioral patterns and should not be used for the sole purpose of special education identification or for psychological evaluations. Once universal screening measures are completed, data obtained from the measures are used to identify students who may be at-risk for academic and behavior difficulties (Eklund et al., 2009). The main goal of implementing universal screening measures in the classroom is to identify students who may benefit from Tier 2 or Tier 3 supports, as they may be experiencing heightened symptoms of internalizing and externalizing problems. The benefit of using universal screening measures, when compared to multiple-gate screening measures, is that it takes pressure off teachers to identify students. Teachers are not trained in psychopathology and should not be expected to be knowledgeable of the various topographies of emotional and behavioral disorders (Lane et al., 2012).

There are some drawbacks to using universal screening measures. Completing screening measures for every student in the classroom has the potential to take a substantial amount of time. Identifying screening measures that are technically adequate

and that utilize minimal time (less than two minutes per student) to complete is often a challenge faced by most schools. Researchers have also criticized the lack of adequate sensitivity and specificity in current universal screening measures (Lambert, Epstein, & Cullinan, 2014). This is a significant cause for concern, as it runs the risk of not accurately identifying students who are at-risk. Another consideration is the cost of a universal screening tool. Recently, some universal screening measures have been developed and validated to address the limitation (i.e. Emotional and Behavioral Screener; Cullinan & Epstein, 2013) of classification accuracy. Some of the available screening tools are published through testing companies and must be purchased (i.e. Behavior and Emotional Screening System; Kamphaus & Reynolds, 2007). For larger school districts or school districts that have limited resources, this may not be practical. However, there are some screening measures that are available at no cost. A review of the most common and current universal screening measures is explored and discussed below.

Social Skills Improvement System: Performance Screening Guide (SSiS-PSG)

The SSiS-PSG is one part to a comprehensive system of assessments, which also includes an intervention program (Lane et al., 2012). Each one can be used individually or together to implement a complete system of screening, monitoring, and targeted interventions for students with behavioral problems (Lane et al., 2012). The SSiS-PSG has three different versions: preschool, elementary, and secondary. Each version is criterion referenced (Elliot & Gresham, 2007). The SSiS-PSG screens for the following domains: prosocial behavior, motivation to learn, math skills, and reading skills.

Operational definitions are provided for each of the domains, as well as key behaviors or skills that would qualify a student to receive a specific rating score. The range of scores for the preschool version is 1 (elevated risk), 2 (moderate risk), and 3 or 4 (adequate performance). As for elementary and secondary, ratings are 1 (experiencing significant difficulty), 2 or 3 (moderate difficulties), and 4 or 5 (adequate performance). The SSiS-PSG is criterion referenced, which means the authors of the assessment tool have pre-determined acceptable levels of performance compared to inadequate levels of performance. Teachers rate every student in their classroom on the four domains, using a color coding system. The SSiS- PSG can be completed for an entire class in approximately 30 minutes (Lane et al., 2012). Materials can be purchased online from Pearson and cost ranges from $19.60 for a packet of four forms (preschool) to $49.05 for a packet of 10 forms (elementary and secondary). Each form is able to screen an entire class of 25 students. Classrooms with more than 25 students require an additional form.

Overall, research has established the SSiS-PSG as an adequate screening tool, with evidence to suggest the SSiS-PSG is a reliable and valid measure (Lane et al., 2012). During the pilot study, social validity was confirmed by 98% of the teacher participants as they reported items were relevant (Elliot & Gresham, 2007). In the same study, test-retest reliability was explored and results provided strong evidence to suggest that the SSiS-PSG is a screener that consistently identifies students experiencing heightened levels of difficulties in one of the four domains. Interrater agreement has also been established as moderate to strong (Elliot & Gresham, 2007). Research has explored concurrent and criterion validity, but only using other measures from the SSiS

assessment packet and achievement tests. The SSiS-PSG is a universal screening measure which has emerging evidence to suggest adequate reliability and validity. It is easy to score, includes behavior skills as well as academic skills, and results are linked to preventative strategies and interventions (Lane et al., 2012). There are limited number of peer-reviewed studies that have explored additional psychometric properties such as classification accuracy. Furthermore, Lane et al. (2012) discussed how the price may be burdensome to school districts that have limited resources or if classrooms are larger than 25 students.

Behavior and Emotional Screening System (BESS)

The BESS is a measure that teachers and parents complete to determine the emotional and behavioral functioning of students between the ages of 3 to 18 (Kamphaus & Reynolds, 2007). The items on the scale are broken into four separate dimensions: Adaptive Skills, Internalizing Problems, Externalizing Behaviors, and School Problems. Three different rating forms are available in English and Spanish, which include parent, teacher, and student. Depending on the form, there are 25-30 items. It takes between five to ten minutes to complete, per student. Behaviors are rated on a four-point Likert scale: never, sometimes, often, or almost always. Scoring options are available for hand and computer. Raw scores are transformed into T-scores that are based on normative samples and resemble the 2002 U.S. census. The BESS kit is estimated to cost between $116.90 and $143.15. The manual costs $78.75. A packet of 25 forms cost $30.45 and a packet of 100 forms is $121.80. The materials are available for purchase online from Pearson.

Several studies have investigated the psychometric properties of the BESS, including the test developers and independent researchers (see Kamphaus & Reynolds, 2007; Kamphaus et al., 2007; King, Reschley, & Appleton, 2012). Kamphaus and Renyolds (2007) reported evidence of acceptable internal consistency, test-retest reliability, and inter-rater reliability. Moderate correlations were found with the SSiS Teacher Form scores and student achievement but no significant correlations between student absences and suspensions (Kamphaus & Reynolds, 2007). Diagnostic accuracy was evaluated using the Behavior Assessment System for Children, 2nd Edition (BASC-2; Reynolds & Kamphaus, 2007) as an outcome measure. However, findings from this study are questionable given that 24 of the 27 items come directly from the BASC-2. King et al. (2012) also published their findings that revealed weak correlations between BESS Teacher Form and ODRs, suspensions, and reading ability.

Some limitations to the BESS are worth noting. For example, initial scoring is difficult to complete and scoring software is somewhat expensive (Lane et al., 2012). If the web-based scoring system is purchased by the school district to aide in universal screening practices and data monitoring, significant time and money is required to get the most out of the software. On the other hand, teacher training is still required for hand scoring, to transfer the raw score to the T-score, and to determine level of risk per student. Lane et al. (2012) argued the BESS currently lacks empirical evidence regarding validity. Overall, limited studies have demonstrated the use of screening data linked to the recommended interventions and whether or not the interventions have been successful at remediating the identified problem behaviors.

Student Risk Screening Scale (SSRS)

The SRSS is a teacher-rating scale used to identify students in kindergarten through sixth grade, who may be presenting with antisocial behaviors (Drummond, 1994). A total of seven behaviors are listed and teachers are asked to rate each of their students on a three-point Likert scale (0 = Never, 1 = Rarely, 2 = Occasionally, 3 = Frequently) for all seven behaviors. Behaviors measured including stealing; lying, cheating, sneaking; behavior problem; peer rejection; low academic achievement; negative attitude; and aggressive behavior. A total score is generated by summing up the responses to all seven items, with a minimum total score of 0 to a maximum total score of 21. Drummond (1994) developed the following risk classification system based on the total score obtained: Low (0-3), Moderate (4-8), and High (9-21). Results from the SRSS can facilitate decision-making as to what level of intensity of intervention is warranted (Tier 2 versus Tier 3). Lane et al. (2012) also suggested that the SRSS be used as a way of monitoring risk over time. Teachers are able to complete this measure in approximately 10 to 15 minutes for their entire classroom (Dever et al., 2012). The SRSS is available online for free at http://www.sai-iowa.org/10_%20Behavior%20Screeners.pdf.

Several peer-reviewed studies have explored and validated the psychometric properties of the SRSS (see Lane et al. 2012 for a comprehensive list). The SRSS has demonstrated low to moderate correlation coefficients for internal consistency and test-retest reliability (Lane et al., 2007). Drummond (1994) found that students' scores on the SRSS predicted negative behavior outcomes and academic difficulties 1 ½ to 10

years later. The SRSS demonstrates acceptable convergent validity with the SSBD for externalizing behaviors but not internalizing behaviors (Lane et al., 2007). The SRSS was designed to identify students with externalizing behaviors and not for internalizing behaviors, which should be considered before selecting this screener (Lane et al., 2012). More recently, Menzies and Lane (2012) conducted a study with the SRSS and found that ratings obtained at the beginning of the year were able to predict the number of ODRs a student received at the end of the year, as well as, self-control skills as reported on the SRSS and reading competency at the end of the school year. As this is just a few of the studies available, there is a substantial amount of evidence to suggest the SRSS is reliable and valid universal screening measure to use with elementary students (see Lane et al., 2012 for additional studies). Emerging evidence supports the technical adequacy of the SRSS to be used with secondary students, both middle and high school.

Emotional and Behavioral Screener (EBS)

The EBS was developed by Cullinan and Epstein (2013) in order to universally screen for students who may be presenting with severe emotional and behavior impairment. Specifically, this screening tool identifies those who may be at risk for later identification of an Emotional Disturbance (ED) special education eligibility. The EBS aligns directly with the five eligibility criteria for ED and may be used with students in kindergarten through 12th grade. This screening measure is the first to align with the federal criteria of ED as outline in IDEA (Cullinan & Epstein, 2013). A total of ten items were selected from the Scales of Assessing Emotional Disturbance, Second Edition (SAED-2; Epstein & Cullinan, 2010) to generate the EBS. The SAED-2 is a

nationally normed assessment tool used for the purpose of identifying students for special education and is not appropriate to use as a universal screening measure. Each item is rated on a four-point Likert scale ranging from 0 (not a problem) to 3 (severe problem). A total score is summed from all items. Students who receive an overall rating that exceeds the cut score provided in the manual (based on age) are identified at-risk. The only form available is a teacher rating scale. The complete kit may be purchased online from Pro Ed for $125.00. Additional rating forms may be purchased in a packet of 50 for $35.00 and the decision summary form for $35.00.

The reliability and validity of the EBS was established in the original study conducted by Cullinan and Epstein (2013). In the same study, the authors were also able to establish acceptable construct validity, as the EBS was able to discriminate between groups of students who are more likely to be identified as ED (i.e. males vs. females, older students vs. younger students). In a follow up study, researchers found similar results (Nordess, Epstein, Cullinan, & Pierce, 2014). Lambert et al. (2014) conducted another study to further investigate the classification accuracy of the EBS. Results from this study support the use of EBS as a universal screening measure that accurately discriminates between students who are at-risk for being identified as a student with an ED disability. Lambert et al. (2014) also reported that across all of their findings, the EBS performed "markedly better for younger students compared to older students" (p. 58). As this is considered a relatively new screening measure, there is limited research available and additional research is warranted to continue validating the psychometric properties of the EBS.

Strengths and Difficulties Questionnaire (SDQ)

The SDQ contains 25-items that measure student's maladaptive behaviors, as well as prosocial behaviors (Goodman, 1997). Teacher and parent rating forms are available for students between the ages of 4 -17 and self-report rating forms are available for students between the ages of 11-17. An early-years version is available and appropriate to administer to children ages 2 to 4. All rating scales for the SDQ are free and available to download online at http://www.sdqinfo.org. Items on the SDQ are classified into the following subscales: Conduct Problems, Hyperactivity/Inattention, Emotional Symptoms, Peer Relationship Problems, and Prosocial Behavior. Each subscale contains five of the 25 items. Each item is rated on a three point Likert scale (0 = Not True, 1 = Somewhat True, and 2 = Certainly True). Four of the five scales are summed into a Total Difficulties score, leaving Prosocial Behavior as a separate score that is positively scored. There is an impact scale available and follow up questions to use when administering after the individual receives intervention. The SDQ has been used worldwide and is available in several different languages (Hill & Hughes, 2007).

The SDQ may be used for clinical assessment, evaluating response to intervention, research, epidemiological studies, and universal screening (Goodman, 1997). There are several different scoring options for the SDQ, which include traditional hand scoring, black and white overlays, or online (http://www.sdqinfo.org/py/sdqinfo/c0.py). The website also contains syntax to score using different statistical software programs such as STATA, SAS, SPSS, and R.

Student's level of risk or level of impairment is determined by normative cut scores that are available online.

There are several strengths worth mentioning regarding the SDQ. First, numerous studies have found strong evidence of reliability and validity, along with normative data for six countries (Australia, Britain, Finland, Germany, Sweden, and the United States). The SDQ is available for free and has been translated into 69 different languages (Lane et al., 2012). In addition, Stone, Otten, Engels, Vermulst, and Janssens (2010) conducted a meta-analysis that revealed satisfactory internal consistency, test-retest reliability, and inter-rater agreement for teacher and parent forms across 48 studies (N = 133,223; 4 to 12 year olds). Validity of the SDQ was also explored in the meta-analysis and found 15 out of 18 studies confirmed a five-factor structure (Stone et al., 2010). When used as a screener and predicting the onset of a disorder, the SDQ has acceptable sensitivity and good specificity (Goodman, Ford, Simmons, Gatward, & Meltzer, 2000). When compared to other screening measures such as the SSRS (Gresham & Elliot, 2008), the SDQ does not provide recommended interventions or strategies based on the screening results. Another limitation, teachers have shared concerns regarding time to complete the screener per student (Lane et al., 2012). With 25 items, it takes approximately five minutes per student to complete, a factor that should strongly be considered prior to implementation.

Behavior Screening Checklist III (BSC)

The BSC (Muyskens et al., 2007) is a 12-item checklist designed to rate behaviors on three separate subscales: Classroom Behaviors, Externalizing Behaviors,

and Socialization. A Total score is also generated by summing ratings across all 12 items. The BSC was developed to measure classroom behaviors for students between kindergarten and 8th grade. Muyskens et al. (2007) originally developed the BSC for teachers to quickly rate a range of essential classroom behaviors during a behavior consultation. Each item is on a 5 point Likert rating scale, with 1 representing behaviors that are consistently observed in the classroom and 5 representing classroom behaviors that are rarely observed. Time to complete the BSC per student is approximately two minutes.

The overall level of risk for each student is determined by summing the rating of each item, which generates a raw score (King & Reschly, 2014). In the original study on the BSC a cut score of 36 was determined for identifying students at-risk (Muyskens et al., 2007). A score of 36 or higher on the BSC included the top 5% of rating scores obtained. This percentage would be consistent with students in need of the most intensive preventative interventions (e.g. Tier 3). In a second study conducted with the BSC, King and Reschly (2014) calculated a cut score of 27 to capture the top 20% of their sample. This percentage of students would typically include those who are in need of Tier 2 and Tier 3 supports.

Currently, there are only two published studies that have explored the psychometric properties of the BSC (King & Reschly, 2014; Muyskens et al., 2007). Muyskens et al. (2007) first piloted the BSC with a sample of 22,056 students between kindergarten and 8th grade. Of these participants, 51.1% were male and included the following racial distribution: African American (40%), White American (28%), Hispanic

American (16.6%), Asian American (11.8%), and Native American (3.6%). Sixty-eight percent of the participants received reduced or free lunch. Internal consistency for the piloted sample was strong, with Cronbach's alpha ranging from .92-.95 across all grade levels. Six pairs of third grade teachers were recruited to explore the inter-rater function of the BSC. Inter-rater reliability was statistically significant ($p < 0.01$) across six pairs of teachers, with a sample size of 143 students. Correlation coefficients between pairs of teachers ranged from .66 to .97, with an average inter-rater reliability of .83 (Muyskens et al., 2007).

Muyskens et al. (2007) investigated predictive validity, focusing on change in screening scores from elementary school to middle school. A separate correlation analysis was conducted using kindergarten through 5th grade as one sample (N = 14,335) and grades 6th through 8th as the second sample (N = 6,721). It was hypothesized that as students progressed through school, a greater number of difficulties would be reported. Spearman's Rho was calculated between Fall BSC scores and end of year suspensions, absences, and end of year reading and math district achievement scores. Analyses found statistically significant correlations ($p < 0.001$) between the Fall BSC scores and all criterion variables ranging from 0.19 to 0.51, with stronger correlations in grades 6th through 8th.

In the second study of the BSC, King and Reschly (2014) investigated the concurrent validity of the BSC with the BESS among students in kindergarten through 5th grade. A total of 23 elementary teachers completed the screeners for each of their students. Four-hundred and ninety two participants were included in the sample, of

which 47% were male, 65% were Caucasian, 82% identified as English Language Learner, and 68% received reduced or free lunch. Results of this study found strong and statistically significant correlations between BSC and the BESS (r = .85). In addition, Fall BSC and BESS scores were correlated on a statistically significant basis with spring outcome measures (e.g. ODRs, number of days suspended, oral reading fluency score, absences, and state achievement testing results). However, it should be noted that predictive validity was different for the BSC and BESS. The fall scores on the BESS demonstrated stronger predictive validity for the achievement variables (state testing results and absences), whereas the BSC showed stronger evidence of predictive validity for the behavior outcomes (ODRs and suspensions).

Overall, King and Reschly (2014) recommend that factors such as time, cost of administration, and feasibility should be considered prior to selecting which measure to use for screening purposes. Given that only two studies have been published to date, additional research is needed to understand how the BSC functions in different schools, with different samples, and across different grade levels. Future studies should focus on elementary school students, given the findings were weaker with this age group of students compared to middle school students. Muyskens et al. (2007) also noted the importance of future research to evaluate the classification accuracy (i.e. sensitivity, specificity) to further understand the predictive validity of the BSC.

Data that Inform Universal Screening Practices

Aside from the multigate screening measures and universal screening measures, schools have an abundance of data at their fingertips that are often considered or utilized

to assist in identifying students at-risk. This includes things such as ODRs and attendance. As with any assessment tool, there are pros and cons to these sources of data.

Office Discipline Referrals

Office discipline referrals (ODRs) are one of the most common sources of existing sources of data used by researchers for universal behavioral screening (Bezdek, 2014). According to Sugai, Sprague, Horner, & Walker (2000), ODRs are defined by the following criteria: "a) a student engaged in behavior that violated a rule or social norm in the school, b) the problem behavior was observed or identified by a member of the school staff, and c) administrative staff delivered a consequence through a permanent (written) product that defined the whole event" (p. 96). Typically students will receive an ODR for classroom disruptions, aggressive behaviors, or skipping. Researchers have recommended school districts follow a decision rule or cut score when using ORDs as a universal screener in order to identify students in need of additional behavior support (Bezdek, 2014; Sugai et al., 2000). Students who receive zero or one ODR within an academic school year is considered to be within normal range. These students generally respond appropriately to universal supports. Students who receive two to five ORDs are identified as at-risk and should strongly be considered for more intensive intervention (e.g. Tier 2). Lastly, students who receive six or more ORDs require the most intensive and individualized interventions that are available (e.g. Tier 3; Sugai et al., 2000).

The use of ODRs as a source of screening data is attractive for schools because it does not require a significant amount of time from teachers or school staff. However,

the current literature provides mixed results on the utilization of ORDs and predicting behavior outcomes for elementary students (Pas, Bradshaw, & Mitchell, 2011). Tobin and Sugai (1999) reported small to moderate correlations with later behavior problems, such as oppositional and defiant behavior; drug use; challenging classroom behaviors; and school dropout. In addition, low to moderate convergent validity has been established between ODRs and clinical problems such as social skill deficits, aggression, delinquency, and attention problems (Morgan-D'Atrio, Northrup, LaFleur, & Spera, 1996). However, it is important to keep in mind these studies were conducted with middle and high school students. Pas and colleagues (2011) discussed how much less is known regarding the use of ORDs and predicting future behavior problems for elementary students.

In a more recent study published by Miller et al. (2015), researchers explored the classification accuracy of ODRs, along with the SSiS, and the Direct Behavior Rating-Single Item Scale (Chafouleas, 2011) using the BESS as the criterion measure. There were a total of 1,974 participants enrolled in first through 8th grade. Screening measure data was also collected at fall, winter, and spring time points in the school year. Miller et al. (2015) found that ODRs did not yield statistically significant area under the curve (AUC) results (.49 to .50). The lack of statistically significant findings suggest that ODRs are able to capture students at-risk about 50% of the time, which is no better than chance. Aside from the lack of statistical significance, ODRs also resulted in poor sensitivity (.21 to .36), excellent specificity (.92 to .97), poor positive predictive values (.50 to .58), and good negative predictive values (.84 to .87). These findings published

by Miller et al. (2015) suggest that as ODRs may be an appropriate source of data to monitor system level trends, their findings provide evidence that ODRs may not accurately predict individual student outcomes for elementary students.

School Absences

School absences refer to the number of days a student is not present at school. Current research supports a relationship between school attendance and academic performance, as well as with behavior functioning in the classroom (Freeman et al., 2016). Specifically, students who have higher rates of school absences are at greater risk for a range of risky behaviors such as violence, sexual behavior, teen pregnancy, suicide attempt, as well as drug and alcohol use (Kearney, 2008). Negative academic outcomes have also been highly correlated with school absences, which include poor grades (Tanner-Smith & Wilson, 2013), failing performance on state standardized assessments (Epstein & Sheldon, 2002), and even school dropout (Rumberger, 2011). In addition, prolonged absenteeism is often strongly associated with physical and/or psychiatric health problems (Kearney, 2008), such as internalizing disorders (Wood et al., 2012). Longitudinal studies also found for students who dropped out of high school, school attendance during first grade was identified as a significant predictor variable (Alexander Entwisle, & Horsey, 1997). Students who dropped out of high school had an average of 16 absences during the first grade, compared to those who graduated had an average of 10. Each additional absence accounted for a 5% increase in likelihood of later school dropout (Alexander et al., 1997).

Despite the significant findings between student absenteeism and academic outcomes, the samples used in all of the referenced studies included middle to high school students. The results may not be generalizable to elementary students. Carroll (2013) argues how current research regarding student absences and poor emotional and behavioral outcomes is limited when it comes to elementary students. Carroll (2013) conducted a literature review of studies that focused on this specific age group of students and found six published studies. Of the studies identified, one study found small, statistical significance between student absences and behavior problems (Gottfried, 2009). Schools collect attendance on a daily basis, therefore it is a source of data that is readily available to schools. Given the lack of research available to facilitate the understanding of the relationship between elementary student absences and emotional and behavior problems, further research is warranted.

Statement of Problem

There are several different types of universal screening measures available for schools to use in order to identify students at-risk for experiencing a range of emotional and behavioral problems (i.e. BESS, SSBD, SDQ). Universal screening measures for behavior have continued to develop as a response to the push towards early identification of students who may be at-risk for emotional and behavioral difficulties. Each of the screening measures discussed above have emerging or established evidence of reliability and validity. However, very few studies have explored classification accuracy by investigating the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Other concerns related to universal screening measures are how

many of them may not meet the feasibility characteristic put forth by Glover and Albers (2007). Some of the universal screening measures have a substantial cost to purchase, time to score is excessive, and are often difficult to maintain over time. Of all the universal screening measures, the BSC currently has emerging evidence supporting the psychometric properties. In order for the BSC to continue to be used as a universal screening measure in the classrooms, additional research is warranted. Furthermore, it is imperative that research investigate additional evidence of reliability and validity, as well as the classification accuracy of the BSC to determine how well the screening measure is able to differentiate between students who are at-risk compared to students who are not at risk.

CHAPTER III

METHODS

Participants

Participants were selected from a public school district located in southeast Texas. The school district consists of 12,822 students across 18 campuses. A total of three campuses were selected by school administrators to participate in the study.

Teacher Participants

A total of 11 general education teachers consented to participate (100% consent return rate). One hundred percent were female. Teacher participants were White/Caucasian (63.6%), American Indian/Alaskan Native (9.1%), and Hispanic/Latino (9.1%). The remaining 18.2% of teacher participants did not report their ethnicity. The average number of years teaching was reported as 10.78 (n = 9), ranging from 1 year to 27 years. Of the teachers who reported, 81.8% of them were certified to teach elementary school and 45.5% were certified to teach early childhood. One teacher was certified to teach English Language Learner students and taught the bilingual kindergarten class that participated in the study. One teacher was special education certified. There were a total of nine teachers who earned a bachelor's degree (81.8%). The remaining teachers did not report highest degree earned. A total of 45.5% of teacher participants completed their educator preparatory program at a university, compared to 36.4% who completed an alternative certification program.

Student Participants

A total of 96 kindergarten students were included in the study (35% consent return rate). Fifty-two percent were male (*n* = 50) and 47.9% were female (*n* = 46). At time of enrollment, the average age of student participant was 5.31 years. The student participants included 45.8% Caucasian (*n* = 44), 35.4% Hispanic/Latino (*n* = 34), 8.3% Black/African American (*n* = 8), 9.4% were identified as having more than one ethnicity (*n* = 9), and 1% American Indian/Alaska Native (*n* = 1). Information regarding identified disabilities and English Language Learner status was not collected.

Participant Recruitment

Teacher and student participants were recruited and enrolled at the start of the 2015-2016 school year. The principal investigator met with kindergarten teachers to first recruit as potential participants. The purpose of this meeting was to provide an overview of the study, address any questions or concerns, and collect teacher consent forms. Upon teacher enrollment, they were given parental consent forms to send home with every student in their kindergarten classroom. There were approximately 25 student in each class.

Teachers sent home parent consent forms beginning on September 10th, 2015 and continued to collect parent consent forms until October 19th, 2015. Any student who returned a signed parent consent form was included in the study as a student participant. Students were excluded from participation if they failed to have a consent form signed by a parent or legal guardian upon the first time point of data collection. Teacher participants followed up with all parents who did not submit a signed consent form by

sending a second consent form home with their student during the last week of collecting consent forms. Each teacher participant received $3 per student with a signed consent form at the end of the study for reimbursement of their time. See Table 2 for the breakdown of student participant per class.

Table 2

Number of Student Participants per Class

Teacher	Number of Student Participants
A	3
B	11
C	8
D	9
E	12
F	7
G	7
H	10
I	15
J	8
K	6
Total	96

Measures

Behavior Screening Checklist-III (BSC)

The BSC (Muyskens et al., 2007) is a 12-item, universal behavior screening measure used to determine the degree to which a student participant was experiencing behavior challenges in the classroom as reported by a teacher. To complete the BSC, teachers were asked to rate the degree to which a student exhibits specific classroom

behaviors. These behaviors are rated on a Likert scale ranging from 1 (exhibits identified behavior) to 5 (does not exhibit identified behavior). Lower scores obtained on the BSC suggest the student displays appropriate classroom behaviors and potentially indicates less of a risk for behavior concerns. The BSC yields three subscale scores (Classroom Behaviors, Externalizing Behaviors, and Socialization) and a Total score. Each subscale score is generated by summing the rating on the inclusive items, which is four items per subscale. A Total score is calculated by summing ratings across all 12 items. For the current study's sample, internal consistency for the Fall BSC ($\alpha = .94$), Winter BSC ($\alpha = .90$), and Spring BSC ($\alpha = .94$) were all acceptable (Cronbach, 1951).

Strengths and Difficulties Questionnaire (SDQ)

The SDQ is a 25-item emotional and behavioral screening measure that was used as the criterion (Goodman, 1997). Teachers were asked to rate student's behaviors observed in the classroom on a 3-point Likert scale (Not True = 0, Somewhat True = 1, and Certainly True = 2). The SDQ ratings are combined to yield five subscales and a total score. The five subscales are: Emotional Symptoms, Conduct Problems, Hyperactivity/Inattention, Peer Relationship Problems, and Prosocial Behavior. A Total Difficulties score is generated using all of the subscales with the exception of the Prosocial Behavior subscale. Ratings provided on the SDQ were scored using syntax (available online) for SPSS to generate the SDQ scores. Normative data and recommendations on how to classify scores are available online (http://www.sdqinfo.org/g0.html).

Subscales and total scores can be categorized into one of the following descriptive categories: Normal, Borderline, and Abnormal (see Table 3).

Table 3

Risk Categories and Raw Score Ranges for Interpreting Ratings on the Strengths and Difficulties Questionnaire

Scale/Subscale	**"Normal"**	**"Borderline"**	**"Abnormal"**
Total Difficulties	0 – 11	12 – 15	16 – 40
Emotional Symptoms	0 – 4	5	6 – 10
Conduct Problems	0 – 2	3	4 – 10
Hyperactivity/Inattention	0 – 5	6	7 – 10
Peer Relationship Problems	0 – 3	4	5 – 10
Prosocial Behavior	6 – 10	5	0 – 4

An Abnormal score may be used to identify students who are experiencing clinically significant levels of mental health dysfunction. Those who receive a score within the Borderline range may be experiencing symptoms related to emotional and behavioral difficulties, but not of clinical significance. The Borderline score may be helpful for identifying students at-risk for emotional and/or behavior challenges experienced in the classroom and who may be at-risk for later onset of clinical diagnoses. For the purpose of this study, the Borderline score of 12 or higher was used as the cut-score for the SDQ in all analyses. For the present study yielded internal consistency for the Fall SDQ ($\alpha = .73$), Winter SDQ ($\alpha = .74$), and Spring SDQ ($\alpha = .72$) were all acceptable (Cronbach, 1951).

Office Discipline Referrals (ODRs)

Office discipline referrals indicate the number of times a student was sent to the office during the school year for a conduct-related infraction. ODRs were collected from the Public Education Information Management System (PIEMS) by a school administrator. Data were reported for each student participant and as an aggregate sum for each phase of data collection.

School Absences

School absences refers to the total number of days a student was not present for at least half of the school day. Student absences were also collected from the PIEMS. There were a total of 187 instructional days that a student was required to attend for the 2015-2016 school year. Student absences was reported for each phase of data collection as an aggregate number.

Procedures

Data Collection

Data collection began after consent forms were signed and collected from students' parents. Data collection occurred during the fall (10/23/2015 through 10/30/2015), winter (2/5/2016 through 2/12/2016), and spring (5/20/2016 through 5/27/2016) of the 2015-2016 school year. At each data collection phase, teacher participants completed two universal screening measures (the BSC and SDQ) for each student participant in their respective classrooms. Teachers were given one week to complete the BSC and the SDQ. District administrators provided the total number of office discipline referrals and total number of accrued absences at each data collection.

CHAPTER IV

RESULTS

Overview of Analysis Plan

Research Question #1

How consistent are the BSC Total scores across administration with a sample of kindergarten students? Bivariate correlations were used to analyze the stability and consistency of BSC Total scores over the course of an academic year. Spearman's Rho was selected because of the ordinal nature of the independent (BSC scores) and dependent variables (SDQ scores). All assumptions for using Spearman's Rho were met.

Research Question #2

To what degree does each administration of the BSC demonstrate convergent validity with the SDQ? Convergent validity was investigated using bivariate correlation analysis to understand the relationship between BSC Total scores and SDQ Total Difficulties across three different time points in the school year (e.g. Fall, Winter, Spring). All assumptions for using Spearman's Rho were met.

Research Question #3

To what degree does the Fall BSC and Winter BSC exhibit predictive validity with the SDQ? Predictive validity was investigated using bivariate correlation analyses to investigate the relationship between Fall BSC and Winter BSC Total scores with Spring SDQ Total Difficulties. All assumptions for using Spearman's Rho were met.

Research Question #4

What proportion of variance in Spring SDQ Total Difficulties scores is accounted by BSC Total scores, number of absences, and ODRs? Three separate multiple regression models were performed. The fall model used Fall ODRs, Fall absences, and Fall BSC Total to predict Spring SDQ Total Difficulties. The winter model used Winter ODRs, Winter absences, and Winter BSC Total scores to predict Spring SDQ Total Difficulties. The spring model used Spring ODRs, Spring absences, and Spring BSC Total scores to predict Spring SDQ Total Difficulties. Each model was generated using the enter method. All assumptions for conducting a multiple regression model were met.

Research Question #5

Does the BSC Total demonstrate acceptable classification accuracy in identifying students at-risk for challenging behaviors? A receiving operating characteristic (ROC) curve analysis was conducted to answer this research question. A cut score for the SDQ Total Difficulties was generated using the recommended scores for "Borderline." Any student who received a SDQ Total Difficulties score equal to or greater than 12, was recoded into a one (e.g. at-risk). Any student who received a SDQ Total Difficulties score equal to or less than 11, was recoded into a 0 (e.g. not at-risk). In using a ROC curve analysis, various cut-scores from the BSC were generated by plotting the true positive rate (sensitivity) in conjunction with the true negative rate (1 - specificity).

Research Question #6

Does the BSC Total (Fall, Winter, and Spring) demonstrate stronger classification accuracy in identifying students at-risk compared to ODRs and student

absences? A ROC curve analysis was conducted to answer this research question. The same cut-score generated for the SDQ in research question five was used in this analysis. Only variables that were identified as statistically significant predictors were used in this analyses.

Descriptive Statistics

Analyses were conducted using IBM SPSS Statistics (version 24.0). All variables were screened for normality and outliers. See Table 4 through 6 for a summary of descriptive statistics across all predictor and outcome variables. A review of skewness and kurtosis statistics indicated a normal distribution across all variables, except for ODRs. Most student participants had zero office discipline referrals across the school year, which produced a positively skewed distribution. The three most common types of transformations (square root, log, and inverse) were attempted with each of the ODR variables. However, none of the transformations improved the distribution, nor did it produce a distribution that approached normality. Therefore, the ODRs were left as the original data reported by the PIEMS for all subsequent data analyses. All correlation coefficients were interpreted using Cohen's (1988) recommendations of .10 as weak, .30 as moderate, and .50 as strong.

Table 4

Fall Predictor Variable Descriptive Statistics

Variable	*M*	*SD*	Min.	Max.	Skewness	Kurtosis
BSC						
Classroom Behavior	8.14	3.93	4	20	0.93	0.27
Externalizing Behaviors	6.85	3.60	4	18	1.24	0.44
Socialization	6.77	3.04	4	18	1.29	1.55
Total	21.76	9.74	12	52	1.18	0.77
SDQ						
Emotional Symptoms	1.33	2.06	0	10	2.02	4.25
Conduct Problems	1.41	2.05	0	8	1.52	1.47
Hyperactivity/Inattention	3.86	3.45	0	10	0.37	-1.21
Peer Relationship Problems	1.36	1.61	0	7	1.26	1.51
Prosocial Behavior	8.01	2.50	1	10	-1.08	0.05
Total Difficulties	7.97	7.07	0	29	0.81	-0.16
ODRs	0.08	0.45	0	4	7.39	61.38
Absences	1.63	1.93	0	8	1.47	1.70

Note. $N = 96$. BSC = Behavior Screening Checklist; SDQ = Strengths and Difficulties Questionnaire; ODRs = office discipline referrals.

Table 5

Winter Predictive Variable Descriptive Statistics

Variable	*N*	*M*	*SD*	Min.	Max.	Skewness	Kurtosis
BSC							
Classroom Behavior	87	7.34	3.75		19	1.52	1.94
Externalizing Behaviors	88	6.80	3.20	4	16	0.97	0.01
Socialization	90	6.29	2.71	4	16	1.60	2.61
Total	86	20.43	8.20	12	41	0.93	-0.06
SDQ							
Emotional Symptoms	90	1.46	1.70	0	7	1.23	0.91
Conduct Problems	90	1.23	1.81	0	8	1.71	2.66
Hyperactivity/Inattention	90	3.11	2.97	0	10	0.67	-0.51
Peer Relationship Problems	90	1.44	1.60	0	6	1.10	0.37
Prosocial Behavior	90	8.60	1.89	2	10	-1.26	0.73
Total Difficulties	90	7.24	5.93	0	24	0.93	0.18
ODRs	96	0.21	1.44	0	14	9.39	90.46
Absences	96	4.00	3.96	0	20	1.85	4.37

Note. BSC = Behavior Screening Checklist; SDQ = Strengths and Difficulties Questionnaire; ODRs = office discipline referrals.

Table 6

Spring Predictive Variable Descriptive Statistics

Variable	*N*	*M*	*SD*	Min.	Max.	Skewness	Kurtosis
BSC							
Classroom Behavior	91	8.13	4.26	4	20	0.91	0.07
Externalizing Behaviors	91	7.33	4.12	4	20	1.30	1.01
Socialization	91	6.98	3.56	4	20	1.52	2.76
Total	91	22.44	10.64	12	55	1.02	0.32
SDQ							
Emotional Symptoms	91	1.70	1.91	0	8	1.22	1.14
Conduct Problems	91	1.66	2.54	0	10	1.57	1.62
Hyperactivity/Inattention	91	3.69	3.47	0	10	0.53	-1.06
Peer Relationship Problems	91	1.37	1.57	0	7	1.35	1.70
Prosocial Behavior	91	7.69	2.66	2	10	-0.95	-0.20
Total Difficulties	91	8.43	7.23	0	33	0.93	0.64
ODRs	96	0.35	1.90	0	18	8.70	80.78
Absences	96	7.34	6.69	0	37	1.87	4.62

Note. BSC = Behavior Screening Checklist; SDQ = Strengths and Difficulties Questionnaire; ODRs = office discipline referrals.

Bivariate Correlation Analyses

Research Question #1

How consistent are BSC Total scores across administration with a sample of kindergarten students? Bivariate correlation analyses were conducted in order to explore the relationship between Fall, Winter, and Spring BSC administrations. See Table 7 for a summary of bivariate correlation coefficients. Results from bivariate correlation analyses yielded positive and statistically significant ($p < .001$) coefficients among all BSC Total scores. There was a positive and strong correlation between Fall BSC Total and Winter BSC Total ($r = .70$); between Winter BSC Total and Spring BSC Total ($r = .74$); and between Fall BSC Total and Spring BSC Total ($r = .71$). With regard to the BSC subscales, bivariate correlation analysis yielded positive and statistically significant ($p < .001$) coefficients. Classroom Behavior yielded strong, positive correlation coefficients (rs = .54 to .93). Externalizing Behaviors subscales resulted in moderate to strong correlation coefficients (rs = .46 to .90). Socialization subscales also produced moderate to strong correlation coefficients (rs = .39 to .85).

Research Question #2

To what degree does each administration of the BSC demonstrate convergent validity with the SDQ? Bivariate correlation analyses were conducted in order to explore the relationship between paired administrations of the BSC and the SDQ. See Table 8 for a summary of bivariate correlation coefficients for Fall BSC and Fall SDQ. See Table 9 for a summary of bivariate correlation coefficients for Winter BSC and Winter SDQ. See Table 10 for correlation coefficients for Spring BSC and Spring SDQ.

Table 7

Spearman's Rho Correlations Between Fall, Winter, and Spring BSC Scores

		1	2	3	4	5	6	7	8	9	10	11	12
	Fall BSC												
1	Classroom Behavior	1											
2	Externalizing Behavior	.79**	1										
3	Socialization	.67**	.71**	1									
4	Total	.93**	.90**	.85**	1								
	Winter BSC												
5	Classroom Behavior	.72**	.57**	.39**	.64**	1							
6	Externalizing Behavior	.63**	.71**	.51**	.67**	.75**	1						
7	Socialization	.55**	.46**	.49**	.55**	.60**	.63**	1					
8	Total	.72**	.63**	.52**	.70**	.90**	.88**	.84**	1				
	Spring BSC												
9	Classroom Behavior	.73**	.62**	.44**	.68**	.74**	.63**	.48**	.69**	1			
10	Externalizing Behavior	.59**	.66**	.48**	.63**	.60**	.73**	.48**	.68**	.78**	1		
11	Socialization	.54**	.51**	.55**	.58**	.53**	.54**	.53**	.61**	.63**	.70**	1	
12	Total	.70**	.67**	.53**	.71**	.72**	.71**	.54**	.74**	.93**	.91**	.82**	1

Note. BSC = Behavior Screening Checklist. ** $p < .001$.

Table 8

Spearman's Rho Bivariate Correlations between Fall BSC Scores and Fall SDQ Scores

		1	2	3	4	5	6	7	8	9	10
Fall BSC											
1	Classroom Behavior	1									
2	Externalizing Behavior	.79**	1								
3	Socialization	.67**	.71**	1							
4	Total	.93**	.90**	.85**	1						
Fall SDQ											
5	Emotional Symptoms	.19	.17	.35**	.27**	1					
6	Conduct Problems	.60**	.79**	.68**	.74**	.25*	1				
7	Hyperactivity/Inattention	.85**	.80**	.63**	.86**	.29**	.70**	1			
8	Peer Relationship Problems	.54**	.46**	.64**	.60**	.54**	.48**	.49**	1		
9	Prosocial Behavior	-.73**	-.82**	-.76**	-.83**	-.15	-.75**	-.67**	-.54**	1	
10	Total Difficulties	.78**	.74**	.74**	.85**	.61**	.75**	.88**	.75**	-.66**	1

Note. BSC = Behavior Screening Checklist; SDQ = Strengths and Difficulties Questionnaire. ** $p < .001$.

Table 9

Spearman's Rho Bivariate Correlations between Winter BSC Scores and Winter SDQ Scores

		1	2	3	4	5	6	7	8	9	10
Winter BSC											
1	Classroom Behavior	1									
2	Externalizing Behavior	.75**	1								
3	Socialization	.60**	.63**	1							
4	Total	.90**	.88**	.84**	1						
Winter SDQ											
5	Emotional Symptoms	.18	.28**	.34**	.29**	1					
6	Conduct Problems	.52**	.68**	.52**	.63**	.13	1				
7	Hyperactivity/Inattention	.76**	.69**	.55**	.76**	.22**	.51**	1			
8	Peer Relationship Problems	.40**	.52**	.42**	.47**	.38**	.47**	.40**	1		
9	Prosocial Behavior	-.54**	-.65**	-.47**	-.60**	-.01	-.58**	-.46**	-.34**	1	
10	Total Difficulties	.69**	.77**	.65**	.78**	.55**	.68**	.84**	.70**	-.48**	1

Note. BSC = Behavior Screening Checklist; SDQ = Strength and Difficulties Questionnaire. ** $p < .001$.

Table 10

Spearman's Rho Bivariate Correlations between Spring BSC Scores and Spring SDQ Scores

		1	2	3	4	5	6	7	8	9	10
Spring BSC											
1	Classroom Behavior	1									
2	Externalizing Behavior	.78**	1								
3	Socialization	.63**	.70**	1							
4	Total	.93**	.91**	.82**	1						
Spring SDQ											
5	Emotional Symptoms	.16	.12	.31**	.24**	1					
6	Conduct Problems	.74**	.82**	.59**	.80**	.14	1				
7	Hyperactivity/Inattention	.79**	.77**	.59**	.82**	.21**	.72**	1			
8	Peer Relationship Problems	.43**	.51**	.46**	.51**	.40**	.48**	.41**	1		
9	Prosocial Behavior	-.68**	-.79**	-.62**	-.76**	-.13	-.79**	-.71**	-.53**	1	
10	Total	.75**	.78**	.66**	.83**	.51**	.80**	.88**	.69**	-.74**	1

Note. BSC = Behavior Screening Checklist; SDQ = Strength and Difficulties Questionnaire. ** $p < .001$.

Fall BSC and Fall SDQ Convergent Validity

Results from bivariate analysis produced a statistically significant correlation coefficient ($p < .001$) between Fall BSC Total and Fall SDQ Total Difficulties ($r = .85$). Based on these results, it is suggested that ratings obtained on the BSC Total were similar to ratings obtained on the SDQ Total Difficulties during the fall administration. Classroom Behavior subscale on the Fall BSC was not significantly correlated with Emotional Symptoms on the Fall SDQ. However, bivariate correlation analyses yielded statistically significant ($p < .001$) coefficients between Fall BSC Classroom Behavior and SDQ Conduct Problems ($r = .60$), SDQ Hyperactivity/Inattention ($r = .85$), SDQ Peer Relationship Problems ($r = .54$), and SDQ Total Difficulties ($r = .78$). All four of these correlation coefficients were positive and strong. An inverse relationship was observed between the Fall BSC Classroom Behavior and Fall SDQ Prosocial Behavior ($r = -.73$). This correlation coefficient was strong and statistically significant ($p < .001$). The inverse relationship was expected given the Prosocial Behavior subscale is positively scored. When students are rated higher on the Prosocial Behavior subscale of the SDQ (display more appropriate social behaviors), students are also rated lower on the Classroom Behavior subscale of the BSC (less problematic behaviors are observed in the classroom).

The Externalizing Behaviors subscale on the Fall BSC was not significantly correlated with the Emotional Symptoms subscale on the Fall SDQ. However, bivariate correlation analysis yielded statistically significant ($p < .001$) coefficients between Fall BSC Externalizing Behaviors subscale and SDQ Conduct Problems ($r = .79$),

Hyperactivity/Inattention (r = .80), Peer Relationship Problems (r = .46), and SDQ Total Difficulties (r = .74). These four correlation coefficients were all strong. An inverse relationship was observed between the Fall BSC Externalizing Behaviors subscale and the Fall SDQ Prosocial Behavior subscale (r = -.82). This correlation coefficient was strong and statistically significant ($p < .001$). Bivariate correlation analyses yielded statistically significant ($p < .001$) coefficients between the Fall BSC Socialization subscale and all five subscales on the Fall SDQ, as well as the Fall SDQ Total Difficulties. The correlation coefficients ranged from moderate to strong (rs = .35 to .74). All of the correlation coefficients were positive, except an inverse relationship was observed between the BSC Socialization and SDQ Prosocial Behavior subscale.

Winter BSC and Winter SDQ Convergent Validity

Results from bivariate correlation analysis yielded statistically significant ($p < .001$) correlation coefficients that were strong and positive, between Winter BSC Total and Winter SDQ Total Difficulties (r = .78). Based on these results, it is suggested that ratings obtained from the BSC Total were similar to ratings obtained from the SDQ Total Difficulties during the winter administration.

Classroom Behavior subscale on Winter BSC was not significantly correlated with Emotional Symptoms on the Winter SDQ. However, bivariate correlation analysis yielded statistically significant ($p < .001$) coefficients between Winter BSC Classroom Behavior and SDQ Conduct Problems (r = .52), SDQ Hyperactivity/Inattention (r = .76), SDQ Peer Relationship Problems (r = .39), and SDQ Total Difficulties Score (r = .69). All four of these correlation coefficients were positive, ranging from moderate to strong.

As expected, an inverse relationship was observed between the Winter BSC Classroom Behavior subscale and the Winter SDQ Prosocial Behavior subscale (r = -.54). This correlation coefficient was strong and statistically significant ($p < .001$).

The Externalizing Behaviors subscale on the Winter BSC was statistically and significantly correlated with the SDQ Emotional Symptoms (r = .28), SDQ Conduct Problems (r = .68), SDQ Hyperactivity/Inattention (r = .69), SDQ Peer Relationship Problems (r =.52), and Total Difficulties (r = .77). All five correlation coefficients were positive and ranged from moderate to strong. An inverse relationship was observed between the Winter BSC Externalizing Behaviors subscale and the Winter SDQ Prosocial Behavior (r = -.65). This correlation coefficient was strong and statistically significant ($p < .001$).

Bivariate correlation analyses yielded statistically significant ($p < .001$) coefficients between the Winter BSC Socialization subscale and SDQ Emotional Symptoms (r = .34), SDQ Conduct Problems (r = .52), SDQ Hyperactivity/Inattention (r = .55), SDQ Peer Relationship Problems (r = .42), and SDQ Total Difficulties (r = .65). All of these correlation coefficients were positive. However, an inverse relationship was observed between the Winter BSC Socialization and Winter SDQ Prosocial Behavior subscale (r = -.47). This correlation coefficient was strong and statistically significant ($p < .001$).

Spring BSC and Spring SDQ Convergent Validity

Results from bivariate correlation analysis yielded statistically significant ($p < .001$) coefficients, that were strong and positive, between Spring BSC Total and

Spring SDQ Total Difficulties (r = .83). Based on these results, it is suggested that ratings obtained from the BSC Total were similar to ratings obtained from the SDQ Total Difficulties score during the spring administration.

The Classroom Behavior subscale on the Spring BSC was not significantly correlated with Emotional Symptoms on the Spring SDQ. However, bivariate correlation analysis yielded statistically significant (p < .001) coefficients between Spring BSC Classroom Behavior and SDQ Conduct Problems (r = .74), SDQ Hyperactivity/Inattention (r = .79), SDQ Peer Relationship Problems (r = .43), and SDQ Total Difficulties Score (r = .75). All four of these correlation coefficients were positive, ranging from moderate to strong. As expected, an inverse relationship was observed between the Spring BSC Classroom Behavior subscale and the Spring SDQ Prosocial Behavior (r = -.68). This correlation coefficient was strong and statistically significant (p < .001).

The Externalizing Behaviors subscale on the Spring BSC was not significantly correlated with Emotional Symptoms on the Spring SDQ. Bivariate correlation analysis resulted in statistically significant correlation coefficients between the Spring BSC Externalizing Behaviors subscale and the SDQ Conduct Problems (r = .82), SDQ Hyperactivity/Inattention (r = .77), SDQ Peer Relationship Problems (r = .51), and SDQ Total Difficulties (r = .78). All four correlation coefficients were positive and strong. An inverse relationship was observed between the Spring BSC Externalizing Behaviors subscale and Spring SDQ Prosocial Behavior (r = -.79). This correlation coefficient was strong and statistically significant (p < .001).

Bivariate correlation analyses yielded statistically significant ($p < .001$) coefficients between Spring BSC Socialization and SDQ Emotional Symptoms ($r = .31$), SDQ Conduct Problems ($r = .59$), SDQ Hyperactivity/Inattention ($r = .59$), Peer Relationship Problems ($r = .46$), and SDQ Total Difficulties ($r = .66$). The correlation coefficients ranged from moderate to strong. As expected, an inverse relationship was observed between the Spring BSC Socialization and Spring SDQ Prosocial Behavior subscale ($r = -.62$). This correlation coefficient was strong and statistically significant ($p < .001$).

Research Question #3

To what degree does the Fall BSC and Winter BSC administration exhibit predictive validity with the SDQ? Bivariate correlation analysis were conducted in order to investigate the relationship between Fall BSC Total and Winter BSC Total with Spring SDQ Total Difficulties. See Table 11 for a summary of bivariate correlation coefficients between Fall BSC and Spring SDQ. See Table 12 for a summary of bivariate correlation coefficients between Winter BSC and Spring SDQ.

Results from bivariate correlation analysis yielded a statistically significant ($p < .001$) coefficient between the Fall BSC Total and the Spring SDQ Total Difficulties ($r = .62$). The relationship between the Fall BSC Total and Spring SDQ Total Difficulties was strong and positively correlated. As for the Winter BSC Total and Spring SDQ Total Difficulties, results from bivariate correlation analysis a yielded a statistically significant ($p < 0.001$) correlation coefficient. The correlation between these two variables was strong and positive ($r = .65$). Based on these results, it is suggested that

ratings obtained from the BSC Total during the Fall and Winter administration, may have subtle differences when compared to ratings obtained from the SDQ Total Difficulties score during the Spring administration.

Of the three subscales on the BSC, none of the Fall BSC subscales were significantly correlated with the Spring SDQ Emotional Symptoms subscales. All other subscales yielded statistically significant correlations between the Fall BSC subscales and the Spring SDQ subscales ($p < .001$). Statistically significant correlation coefficients ranged from moderate to strong ($rs = .37$ to $.70$). Of the Winter BSC subscales, Socialization was the only one that was statistically significant with Spring SDQ Emotional Symptoms subscale ($p < .001$, $r = .27$). All other subscales yielded statistically significant correlations between the Winter BSC subscales and the Spring SDQ subscales ($p < .001$). Statistically significant correlation coefficients ranged from weak to strong ($rs = .28$ to $.67$). All statistically significant correlation coefficients (between Fall BSC and Spring SDQ; between Winter BSC and Spring SDQ) were positive, except for those between the BSC subscales and the Spring SDQ Prosocial subscale.

Table 11

Spearman's Rho Bivariate Correlations Between Fall BSC Scores and Spring SDQ Scores

		1	2	3	4	5	6	7	8	9	10
Fall BSC											
1	Classroom Behavior	1									
2	Externalizing Behavior	.77**	1								
3	Socialization	.67**	.70**	1							
4	Total	.93**	.90**	.85**	1						
Spring SDQ											
5	Emotional Symptoms	.06	-.09	.08	.02	1					
6	Conduct Problems	.59**	.63**	.49**	.63**	.14	1				
7	Hyperactivity/Inattention	.68**	.65**	.42**	.66**	.21**	.72**	1			
8	Peer Relationship Problems	.37**	.40**	.43**	.42**	.40**	.48**	.41**	1		
9	Prosocial Behavior	-.54**	-.62**	-.53**	-.63**	-.13	-.79**	-.71**	-.53**	1	
10	Total Difficulties	.62**	.57**	.47**	.62**	.50**	.80**	.88**	.69**	-.74**	1

Note. BSC = Behavior Screening Checklist; SDQ = Strengths and Difficulties Questionnaire. ** $p < .001$.

Table 12

Spearman's Rho Bivariate Correlations between Winter BSC Scores and Spring SDQ Scores

		1	2	3	4	5	6	7	8	9	10
Winter BSC											
1	Classroom Behavior	1									
2	Externalizing Behavior	.75**	1								
3	Socialization	.60**	.63**	1							
4	Total	.90**	.88**	.84**	1						
Spring SDQ											
5	Emotional Symptoms	.08	.07	.27**	.18	1					
6	Conduct Problems	.46**	.62**	.47**	.56**	.14	1				
7	Hyperactivity/Inattention	.66**	.67**	.49**	.68**	.21**	.72**	1			
8	Peer Relationship Problems	.28**	.41**	.41**	.39**	.40**	.48**	.41**	1		
9	Prosocial Behavior	-.49**	-.63**	-.51**	-.58**	-.13	-.79**	-.71**	-.53**	1	
10	Total Difficulties	.55**	.65**	.55**	.65**	.51**	.80**	.88**	.69**	-.74**	1

Note. BSC = Behavior Screening Checklist; SDQ = Strengths and Difficulties Questionnaire. ** $p < .001$.

Multiple Regression Analysis

Research Question #4

What proportion of variance in Spring SDQ Total Difficulties scores is accounted for by BSC scores, number of absences, and ODRs? Three multiple regression models were performed between end of year SDQ Total Difficulties as the dependent variable and BSC Total, number of absences, and number of office discipline referrals as independent variables. Pairwise correlation coefficients of all variables included in the regression models are reported in Table 13. The first model included student's Fall BSC Total scores, number of absences in the fall, and the number of office discipline referrals accrued in the fall (see Table 14). The second model included students' Winter BSC Total scores, number of absences in the winter, and the number of office discipline referrals accrued through the school year up to the winter data collection (see Table 14). The third model included students' Spring BSC Total scores, absences in the spring, and the number of office discipline referrals accrued through the school year up to the spring data collection point (see Table 14).

For all three regression models, assumptions were verified. An analysis of standard residuals was carried out, which revealed no outliers (Standard Residual Minimum $\leq$ -2.67 and $\geq$ -2.05, Standard Residual Maximum $\leq$ 3.09 and $\geq$ 2.65). Assumptions of collinearity were met and results indicated multicollinearity was not a concern (Tolerance $\leq$ 0.99 and $\geq$ 0.83, VIF $\leq$ 1.21 and $\geq$ 1.01). The data met assumption of independent errors (Durbin-Watson value $\leq$ 2.02 and $\geq$ 1.82). The histogram of standardized residuals exhibited approximately normally distributed errors, as did the

normal P-P plot of standardized residuals. The scatterplot of standardized predicted values showed the data met assumptions of homogeneity of variance and linearity. The data also met the assumption of non-zero variances (Variance $\leq 67.31 \geq 0.20$). For all three regression models, the unstandardized regression coefficients, the standardized regression coefficients, R^2, and adjusted R^2 are reported in Table 14.

Table 13

Pairwise Correlation Coefficients of Variables Included in the Regression Analyses

	Spring SDQ	Fall ODRs	Fall Absences
Fall ODRs	.37**	1	
Fall Absences	.05	.05	1
Fall BSC	.35**	.35**	-.01
	Spring SDQ	**Winter ODRs**	**Winter Absences**
Winter ODRs	.35**	1	
Winter Absences	.01	-.05	1
Winter BSC	.71**	.23*	.71**
	Spring SDQ	**Spring ODRs**	**Spring Absences**
Spring ODRs	.42**	1	
Spring Absences	.03	-.05	1
Spring BSC	.86**	.40**	.08

Note. ODRs = office discipline referrals; BSC = Behavior Screening Checklist; SDQ = Strengths and Difficulties Questionnaire. ***p* < .001.

Fall Model

The fall regression model was significantly different from zero, $F(3, 87) = 30.36$, $p < .001$. The adjusted R^2 value of .50 indicates that 50% of the variability in year-end SDQ Total Difficulties scores was predicted by Fall BSC scores, Fall absences, and Fall ODRs. However, Fall BSC Total was the only statistically significant predictor variable in the model ($p < .001$). These findings indicate students who received higher teacher ratings on the Fall BSC Total also received higher ratings on the Spring SDQ Total Difficulties. Absences and ODRs that had accrued by the fall data collection did not explain additional variance in Spring SDQ scores.

Winter Model

The winter regression model was significantly different from zero, $F(3, 82) = 31.39$, $p < .001$. The adjusted R^2 value of .52 indicates that 52% of the variability in year-end SDQ Total Difficulties was predicted by Winter BSC scores, Winter absences, and Winter ODRs. However, the Winter BSC Total scores and Winter ODRs were the only two statistically significant predictor variables in the model ($p < .001$). These findings indicate students who received higher teacher ratings on the Fall BSC Total and had a greater number of ODRs accrued by the winter data collection, also received higher ratings on Spring SDQ Total Difficulties. Number of absences that accrued by winter data collection did not explain additional variance in Spring SDQ scores.

Spring Model

The spring regression model was significantly different from zero, $F(3, 87) = 86.94$, $p < .001$. The adjusted R^2 value of .74 indicates that 74% of the variability in

year-end SDQ Total Difficulties was predicted by Spring BSC, Spring absences, and Spring ODRs. However, Spring BSC Total was the only statistically significant predictor variable in the model ($p < .001$). These findings indicate students who received higher teacher ratings on the Spring BSC Total also received higher ratings on the Spring SDQ Total Difficulties. Number of absences and ODRs accrued by the winter data collection did not explain additional variance in Spring SDQ scores.

Table 14

Summary of Multiple Regression Analyses for Variables Predicting Year-End SDQ Total Difficulties

Variables	Fall Model			Winter Model			Spring Model		
	b	*SE b*	β	*b*	*SE b*	β	*b*	*SE b*	β
ODRs	2.13	1.25	0.14	0.90	0.37	0.19**	0.32	0.22	0.09
Absences	0.22	0.29	0.06	-0.07	0.13	-0.04	-0.03	0.06	-0.03
BSC Total	0.48	0.06	0.65**	0.59	0.07	0.66**	0.56	0.04	0.83**
R^2		0.51			0.54			0.75	
R^2 *adjusted*		0.50			0.52			0.74	
F		30.36**			31.39**			86.93**	

Note. ODRs = office discipline referrals; BSC = Behavior Screening Checklist; *b* = unstandardized regression coefficient weights; *SE b* = standard error of unstandardized regression coefficient weights; *β* = standardized regression weights. ** $p < .001$.

Classification Accuracy Analyses

The BSC Total scores for Fall, Winter, and Spring, as well as Winter ODRs, were all statistically significant predictors variables of students' Spring SDQ Total Difficulties. As a result, a total of four separate receiving operating characteristic (ROC) curve analyses were conducted to answer research questions number 4 and 5. Due to the fact that Fall ODRs, Spring ODRs, and student absences (at all three data collection phases) were not identified as statistically significant predictor variables in each of the regression models, ROC curve analyses were not conducted using these variables.

A cut score for Spring SDQ Total Difficulties was established using the recommended bandings provided with by the SDQ scoring guidelines (http://www.sdqinfo.org/g0.html). For the purpose of this study, the "Borderline" banding was selected for the cut score, which suggests these student may be at-risk for experiencing a range of emotional and behavioral problems in the classroom as reported by a teacher. Any student who received a score of 12 or higher on the Spring SDQ score was classified as "at-risk" for emotional and behavioral difficulties. Results for each ROC curve analyses and corresponding specificity, sensitivity, positive predictive values (PPV), and negative predictive values (NPV) are provided in Table 15.

Research Question #5

Does the BSC Total score demonstrate acceptable classification accuracy in identifying students at-risk for challenging behaviors? Area under the curve was statistically significant for Fall BSC Total, Winter BSC Total, and Spring BSC Total (asymptotic significance < .001). Area under the curve for the BSC ranged from .84 to

.91, which suggests there was an 84% to 91% likelihood that a randomly selected student in the abnormal group (score of 12 or higher on the Spring SDQ Total Difficulties) would have a higher BSC Total score at any point in the school year compared to a student randomly selected from the normal range. Using Compton, Fuchs, Fuchs, and Bryant (2006) recommended indictors, these results suggest good to excellent discrimination for the BSC. A cut score of 18 on the BSC was selected across all three administrations of the BSC. This was determined by examining all possible cut scores and optimizing the threshold for sensitivity and sensitivity, per each respective cut score. Sensitivity, specificity, positive predictive power, and negative predictive power were calculated for each ROC curve analysis. Based on the cut score of 18, the following students were identified as at risk for year-end emotional and behavioral difficulties: Fall BSC identified 46 of the 91 students, Winter BSC identified 47 of the 86 students, and Spring BSC identified 48 of the 91 students.

When selecting the cut score based on an optimal threshold of sensitivity and specificity, it is recommended to consider clinical and financial factors related to the results of the screening (or diagnostic) measure (Erkel & Pttynama, 1998). Depending on the utility of the screening results, some have recommended that sensitivity fall between 90 to 95% (Jenkins, Hudson, & Johnson, 2007). Ninety-three percent of students who were identified as at-risk on the SDQ at the end of the school year, were also identified as at-risk on the BSC at fall administration. This is excellent in terms of accurately identifying students who may need additional emotional and behavioral support in the classroom and is within the recommendations set for by Jenkins et al.

(2007). The winter administration of the BSC was able to accurately identify 85% of students at risk, which is lower than the fall administration but considered "good" sensitivity. The spring administration of the BSC had the strongest sensitivity, with 96% of students being accurately identified as at-risk.

Across all three administrations of the BSC, specificity was lower. The fall and spring administration accurately identified 67% of students not at-risk. The winter administration of the BSC was the lowest, with only 58% of students accurately identified as not at-risk. Positive predictive value (PPV) was also calculated and is the probability the test result is positive, when in fact the individual is identified with a problem. Based on the results provided, it appears the BSC demonstrated low rates of PPV as they were all below .57. These findings suggest students who received higher ratings on the BSC, were rated lower on the SDQ at the end of the school year. That is, students who were identified at-risk on the BSC were not identified at-risk on the Spring SDQ. Negative predictive value (NPV) is the probability that a person does not have the condition and the screening measure was able to accurately detect its absence. These findings suggest the BSC demonstrates higher rates of NPV, which indicates students who are having less classroom behavior challenges as reported on the BSC, also received more positive behavior ratings on year-end SDQ scores. That is, students who are identified as not at-risk on the BSC were also identified as not at-risk on the SDQ.

Research Question #6

Does the BSC Total score (obtained during fall, winter, or spring) demonstrate stronger classification accuracy in identifying students at-risk compared to student

absences and ODRs? Given that student absences were not identified as a statistically significant predictor variable in any of the regression models, this variable was excluded from ROC curve analyses. BSC Total score for Fall, Winter, and Spring, as well as Winter ODRs were found to be a statistically significant predictor variable for year-end SDQ Total Difficulties, therefore were included in the additional classification accuracy analyses. The area under the curve was not statistically significant for the Winter ODRs (AUC = .59), indicating the number of ODRs in the winter functioned no better than chance at determining year-end behavior risk status. Based on the AUC, the variable that had the strongest classification accuracy was Spring BSC.

Table 15

Classification Accuracy Indicators for the Behavior Screening Checklist Predicting Year-End Strengths and Difficulties Scores

Variable	AUC	Std. Err.	95% CI	Cut Score	Sensitivity	Specificity	PPV	NPV
Fall BSC	.88**	.04	.81 – .96	18.0	.93	.67	.57	.96
Winter BSC	.84**	.05	.74 – .93	18.0	.85	.58	.47	.90
Spring BSC	.91**	.03	.85 – .97	18.0	.96	.67	.56	.98
Winter ODRs	.59	.09	.41 – .78	1.0	.08	1.00	1.00	.87

Note: AUC = area under the curve; Std. Error = standard error; CI = confidence interval for the AUC; PPV = positive predictive value; NPV = negative predictive value. ** $p < .001$.

CHAPTER V

SUMMARY

The goal of this study was to investigate psychometric properties of the BSC with a sample of kindergarten students. Emphasis was placed on kindergarten, as this is often a period of transition from a less structured environment (e.g. preschool, home) to a more structured environment. During this time of transition, students may struggle to adapt as they are learning the norms and behavioral expectations of a classroom, which often involve developing self-regulation, maintaining and sustaining attention for extended periods of time, Prosocial Behavior skills, compliance to adult directives, etc. As research has shown, nearly one half of kindergarten students will experience behavior challenges as they acclimate to the school environment (Rimm-Kaufman, Pianta, & Cox, 2000). Elevated disruptive behavior patterns that emerge early in childhood are likely to remain relatively stable throughout elementary school and into adolescence (Campbell, 1995). Therefore, school administrators and staff must have access to psychometrically sound universal behavior screening measures that are feasible to implement and are accurate when it comes to identifying students at-risk. Research questions explored for this study examined the consistency of BSC scores over time, convergent and predictive validity, as well as classification accuracy. Multiple regression models were also conducted to determine what predictor variables best explain year-end ratings on a criterion measure.

Research Question #1

How consistent are the BSC Total scores across administration with a sample of kindergarten students? The first research question examined stability of BSC scores, when administered over the course of one school year. Specifically, this research question investigated the relationship of scores produced on the Fall administration of the BSC, compared to scores produced from Winter and Spring administration of the BSC. Subscales and Total scores across all three administrations of the BSC were found to have moderate to strong relationship. Teacher ratings on the Fall administration of the BSC were similar to ratings on the Winter and Spring administration of the BSC. Furthermore, findings provide evidence that BSC produces consistent teacher ratings over time.

To date, no research has explored the stability of ratings produced by the BSC. Previous research has focused on establishing reliability with the BSC by using intra-rater and inter-rater reliability. King and Reschly (2014) examined intra-rater relatability by calculating the percentage of students found to be at-risk on the BSC compared to the BESS. For this study, the screening measures were completed one time, after the first nine weeks of school. Muyskens et al. (2007) investigated inter-rater reliability, collecting data at one time point during the school year. It is important to establish the consistency of a universal screening measure, particularly with younger children, because over time one would expect that behaviors change. Given that no other study has focused on exploring the consistency of BSC ratings over the course of a

school year, results are promising for the BSC and add to the current literature on the BSC.

Research Question #2

To what degree does each administration of the BSC demonstrate convergent validity with the SDQ? This research question investigated the degree to which scores on the BSC were correlated with simultaneous administrations of the SDQ. Exploration of convergent validity focused primarily on the BSC Total and the SDQ Total Difficulties. Results indicated a strong and positive relationship between the Total scores on the BSC and the SDQ. These findings were similar across all three administrations. BSC Total scores obtained during the Fall, Winter, and Spring administration were similar to the SDQ Total Difficulties obtained for each respective administration. Theoretically, both measures were developed to measure appropriate classroom behaviors. Based on the results, there is evidence to suggest that total scores on both screeners measure similar constructs. The only study available to date that explored convergent validity of the BSC found strong, positive correlations with the Behavioral and Emotional Screening System (BESS; Kamphaus & Reynolds, 2007). Therefore, results from the present study provide additional evidence and add to the current literature, indicating the BSC may be a valid screening measure for appropriate classroom behaviors.

A majority of BSC subscales were also correlated with SDQ subscales. Statistically significant correlations found between the subscales on the BSC and the SDQ were in the expected direction. For example, the BSC Externalizing Behaviors

subscales were positively correlated with SDQ Hyperactivity/Inattention, SDQ Conduct Problems, and SDQ Peer Relationship Problems subscales. Results suggest students who display higher rates of externalizing behaviors (e.g. fidgeting, impulsivity) as reported on the BSC by teachers, also received higher ratings on the SDQ Hyperactivity/Inattention, Conduct Problems, and Peer Relationship Problems subscales. Items on the BSC Externalizing Behaviors subscale assessed physical and verbal behaviors towards people and property, as well as ability to remain in assigned area. All of these behaviors are commonly associated with Hyperactivity/Inattention. Furthermore, students who have higher rates of externalizing behaviors are at an increased risk of developing poor peer relations and experiencing higher rates of conduct problems in the classroom (e.g. fighting, following directions). Based on the convergent validity analyses, the results contribute to the current literature on the BSC and provide evidence that the BSC is able to detect various externalizing classroom behavior problems as reported by teachers.

The BSC Total and the three subscales were negatively correlated with the SDQ Prosocial Behavior subscale. Higher scores on the SDQ Prosocial Behavior subscale indicate the presence of more positive skills such as sharing with others, kindness, selflessness, and empathy. Therefore, students who exhibit higher rates of Prosocial Behavior skills are more likely to experience lower levels of negative classroom behaviors, are able to pay attention to instruction, are more likely to complete their work on time, and frequently participate in class (BSC Classroom Behaviors subscale). In addition, these students are more likely to have higher rates of positive interactions with

teachers and peers, and may even display greater levels of resiliency as they are able to cope with change. These research findings are not only expected given previous research, but are extremely valuable for intervention development and skills to target for positive change in the classroom.

Results from convergent validity analyses also found no relationship between the BSC Classroom Behavior subscales and the SDQ Emotional Symptoms subscales. Similar results were also found between the BSC Externalizing Behaviors subscales and the SDQ Emotional Symptoms subscales. The lack of correlation among these variables is likely due to the nature of the items included on the BSC. For example, the BSC Classroom Behavior subscale included items related to attention, following directions, completing assignments, and participating in class. The BSC Externalizing Behaviors subscales included items related to physical behavior towards others and school property, as well as verbal behavior and ability to remain in assigned area. Both of these subscales on the BSC (Classroom Behavior and Externalizing Behaviors) measure behaviors not related to those on the SDQ Emotional Symptoms subscales. The Emotional Symptoms subscale on the SDQ includes items that assess a range of internalizing symptoms such as somatization (e.g. headaches, stomachaches), apprehension, sadness, anxiety, and fearfulness. Based on the result, it appears the BSC is not an appropriate measure to assess internalizing behaviors or emotional symptoms.

Research Question #3

To what degree does the Fall BSC and Winter BSC exhibit predictive validity with the SDQ? This research question explored the degree to which scores on the BSC Fall and

Winter BSC administration predict scores obtained on year-end administration of the SDQ. Exploration of predictive validity focused primarily on the BSC Total scores predicting SDQ Total Difficulties scores, with additional examination on the subscale scores. Overall, results indicated a strong and positive relationship between Fall BSC Total and Winter BSC Total with Spring SDQ Total Difficulties. Classroom behaviors as reported on the BSC at the beginning and middle of the school year are similar to classroom behaviors measured on the SDQ at the end of the school year.

Theoretically, both measures were developed to assess the presence of appropriate classroom behaviors. Ratings on the BSC were similar to ratings on the SDQ at the end of the school year. These findings have important implications for behavior screening in the classroom. Results support the BSC as a universal classroom behavior screener that is able to accurately predict behaviors observed in the classroom at the end of the school year. Given the importance of early identification, these findings provide insight for administering universal screening measures early in the school year. Based on these findings, if schools are able to screen and identify students at-risk for developing emotional and behavioral difficulties early in the school year, these students are more likely to receive prevention interventions before the onset of a full blown clinical diagnoses of mental health disorder. When students are identified early, internalizing symptoms and externalizing behaviors tend to be mild, increasing the likelihood of response to intervention (Gresham et al., 2013).

As for the BSC subscales, results revealed a majority of Fall BSC and Winter BSC subscales significantly correlated with Spring SDQ subscales, with the exception of

Emotional Symptoms. Scores obtained on the BSC subscales during the Fall and Winter accurately predicted scores obtained on the SDQ at the end of the year. The Classroom Behavior and Externalizing Behaviors subscales on the BSC had the strongest relationship with the SDQ Conduct Problems and SDQ Hyperactivity/Inattention. These subscales perform slightly better at predicting end of the year scores when compared to the other subscales on the BSC. These findings provide strong evidence to suggest items on BSC Classroom Behavior and BSC Externalizing Behavior subscales are similar to the items on the Conduct Problems and Hyperactivity/Inattention subscales on the SDQ. Overall, these two BSC subscales measure behaviors such as attention, ability to follow directions, work completion, class participation, as well as physical and verbal behavior. These are all behaviors that are similar to those measured on the SDQ Conduct Problems and SDQ Hyperactivity/Inattention subscales. The BSC Socialization subscale produced moderate correlation coefficients with all subscales on the SDQ. The Socialization subscale on the BSC measures behaviors related to interactions with others (peers and adults), self-image, and one's ability to cope with change. These types of behaviors, as measured by the Socialization subscale, are moderately related to the other subscales on the SDQ. For example, if a student has difficulty interacting with peers or teachers (as measured on the Socialization subscale), they are more likely to receive higher ratings on the SDQ Conduct Problems or SDQ Hyperactivity/Inattention subscales.

None of the BSC subscales were found to correlate on a statistically significant basis with the SDQ Emotional Problems subscale, except for the Winter BSC Socialization subscale. The relationship found between the Winter BSC Socialization

and Spring SDQ Emotional Problems subscale was very low (r = .27). Results are similar to the convergent validity research question. Emotional Problems subscale specifically measures internalizing behaviors. Items on the BSC do not appear to align with SDQ Emotional Symptoms. The Emotional Symptoms subscale on the SDQ includes the following behaviors: often complains of headaches; frequently worries; often unhappy, nervous or clingy in new situations; is easily scared or has many fears. These are all considered internalizing symptoms. Furthermore, internalizing symptoms are more overt and hard to pick up. These types of behaviors involve impairment that occur within an individual and involves avoidant behaviors. This is a concern for researchers and practitioners, as younger children do experience a range of internalizing symptoms that are impairing at school and home, all of which warrant clinical attention (In-Albon, 2012). Over the past decade, research has continued to support the idea that internalizing disorders are occurring more frequently in younger children than externalizing disorders (In-Albon, 2012). It is imperative that universal screening measures are developed to identify students at-risk for internalizing disorders such as anxiety or depression. This is a limitation of the BSC that researchers and practitioners need to be aware of and provide additional evidence that it may not be an appropriate screening measure for internalizing disorders.

Overall, the BSC functions exceptionally well when it comes to predicting end of the year SDQ scores. For schools to engage in best practices of universal behavior screening, psychometrically sound measures must be available. Universal screening measures must target a range of behaviors rather than focus on one specific problematic

area. The BSC has 12 items, with four items included in each subscale. Although the Total BSC score should be used to determine overall level of risk for students, the subscale scores may be helpful for identifying specific areas in which a student is struggling. It may also be helpful to use subscale scores when looking at how well (or not well) a student is responding to behavior supports and interventions once they have been identified at risk. In previous studies, predictive validity of the BSC was established with other universal behavior screening measures, as well as school-level data and end of the year achievement (King & Reschly, 2014; Muyskens et al., 2007). The current study extends the current literature on the BSC and provides substantial amount of evidence to support the use of the BSC as predicting end of year academic and behavioral outcomes.

Research Question #4

What proportion of variance in Spring SDQ Total Difficulties scores is accounted by BSC Total scores, number of absences, and ODRs? Using a multiple regression analyses, this question attempted to identify the predictor variables that best account for variance in Spring SDQ scores. The current study was the first to use a multiple regression analyses with the BSC. Three models were conducted to determine what variables would predict Spring SDQ scores. Predictor variables included in all three models were ODRs, absences, and BSC Total scores. The Fall, Winter, and Spring models predicted 50%, 52%, and 74% (using *adjusted* R^2) of the Spring SDQ scores, respectively. Given these results, data collected in the Fall and Winter accounted for a substantial (and approximately the same) amount of variance in the end of the year SDQ

scores. Data collected in the spring, accounted for more variance in year-end SDQ scores.

Although all three regression models were statistically significant, only a few of the predictor variables accounted for a statistically significant amount of variance in Spring SDQ scores. Fall, Winter, and Spring BSC Total ratings accounted for a statistically significant amount of variance. These findings suggest teacher ratings on the BSC predicted behavior ratings as measured by the SDQ at the end of the school year. Fall and Winter BSC ratings accounted for approximately the same amount of change in Spring SDQ scores, with Spring BSC predicting more change. That is, when holding all other predictor variables constant, Spring SDQ scores increased by 0.65 standard units for one standard deviation increase on the Fall BSC ratings. For one standard deviation increase in Winter BSC scores, the SDQ scores increased by 0.66 standard units. Subsequently, Spring SDQ scores increased by 0.83 standard units when Spring BSC scores increased by one standard deviation. Overall, this suggests that higher scores on the BSC, at any time point in the school year, predicted higher ratings in the SDQ at the end of the school year. Based on the multiple regression analysis, the Fall and Winter BSC was able to predict about the same amount of change in the SDQ scores at the end of the year. Spring BSC was able to predict the greatest amount of change in Spring SDQ scores. Winter ODRs was the only other predictor variable that was statistically significant. Number of absences was not a statistically significant predictor variable. These findings offer promising evidence to support using the BSC early in the school year as a universal screening measure for predicting the onset of

behavioral difficulties later in the school year.

Recent research has shown that students who are identified at-risk, will remain at risk over time when they do not receive intervention services (Dowdy et al, 2014; Dever et al. 2015). Therefore, if students are identified at-risk in the fall of a school year, they are more likely to remain at risk at the end of the school year. Results of this multiple regression analysis indicates there are minimal differences between screening students at-risk in the fall compared to the winter. Depending on available resources, it may be appropriate to screen at the start of the school year, that way schools can provide prevention and early intervention services immediately rather than waiting later in the school year. For schools who have limited resources, they may have to wait until later in the school year to screen their students. If at the end of the year, a group of students are identified as at-risk for behavioral problems, these students are more likely the ones who are truly in need for additional behavior supports. If this were a practice schools implemented, they could track the previously identified group of students at-risk during the first part of the school year and have early intervention strategies (Tier 2) available immediately. Screening students at the end of the year may be counterintuitive for researchers and may not be considered as best practice for multi-tiered system of support. However, with only approximately 2% of schools engaging in some type of universal behavioral screening in the classroom (Romer & McIntosh, 2005), this may be a step in the right direction at getting more schools on board with screening practices.

Other studies have investigated the utility of using readily available data, such as ODRs and attendance to inform the screening process. Both variables are practical for

schools to collect and analyze to determine if students are at risk because they are cost effective (e.g. free) and efficient to gather. However, current literature using ORDs (Pas et al., 2011) and attendance (Carroll, 2013) is extremely limited when predicting emotional and behavior outcomes for elementary students. The results from the current study suggest both ODRs and absences are not sufficient at predicting end of the year behavioral problems in the classroom. Although these findings may be an important consideration for moving the field forward, the number of students included in the study who had an excessive number of absences and ODRs was very small. Of the total 96 participants included, only 6.2% had more than two ODRs and only 5% of the participants missed more than 10% (more than 19 days) of instructional days. The lack of statistically significant findings may be due to the small number of students who had excessive absences or ODRs. Although Winter ODRs were identified as a statistically significant predictor variable, it is recommended that schools continue to collect this information and interpret within the context of other sources of data. Previous research is inconsistent when examining the psychometric properties (i.e. predictive validity) of ODRs, particularly with regards to elementary students (Pas et al., 2011). Additional research is necessary to better understand the screening utility of ODRs and student absences.

Research Question #5

Does the BSC Total scores demonstrate acceptable classification accuracy in identifying students at-risk for challenging behaviors? Using a ROC curve analysis, this research question explored how accurate the BSC is at differentiating students who are

truly at risk from experiencing behavior difficulties from those students who are not at risk. The current study was the first to conduct a ROC curve analysis with the BSC. Three separate ROC curve analyses were conducted using Fall BSC, Winter BSC, and Spring BSC.

Area Under the Curve (AUC)

The AUC for the Fall BSC was .88 (or 88%), Winter BSC was .84 (or 84%), and Spring BSC was .91 (or 91%). These findings represent the overall probability that a student who is at-risk would have a higher rating on the BSC compared to those students who are not at-risk. To put in context, the BSC has 84 to 91% probability of capturing students who are at-risk for experiencing classroom behavior problems. Area under the curve represents "the average value of sensitivity for all possible values of specificity" (Park, Goo, & Jo, 2004, p. 13). When interpreting AUC results, values that are greater than 0.90 are identified as excellent, 0.80 to 0.89 are good, 0.70 to 0.79 are fair, and values less than 0.69 are poor (Compton et al., 2006). Based on the results, the BSC has "good" to "excellent" discrimination. These findings provide important implications for schools in moving forward with implementation of universal behavior screeners, and possibly using the BSC. However, there are other diagnostic indicators that must be examined and understood before drawing conclusions about the BSC classification accuracy. As this was the first study to explore classification accuracy with the BSC, the results offer preliminary evidence.

Sensitivity and Specificity

There are a few ways to determine the sensitivity and specificity of a screening or diagnostic measure. For this study, cut scores were selected in order to maximize sensitivity and specificity. Subsequently, sensitivity and specificity often have an inverse relationship. Therefore, when selecting a cut score with a sensitivity rating of .90 or higher, it is important to see how this impacts the specificity rating. Measures that are considered diagnostically accurate, typically will have higher ratings of specificity compared to measures that are less accurate.

Sensitivity is the proportion of students in the sample who were identified as at-risk for behavioral concerns on the SDQ that were also correctly identified as at-risk by the BSC. When selecting a cut score of 18 across all three administrations of the screening measures, sensitivity probabilities indicated that 85 to 96% of participants were identified at-risk on the BSC (Fall, Winter, and Spring) and at-risk on the Spring SDQ. Overall, winter sensitivity was the lowest. This finding may have been due to a higher number of missing data within the BSC for winter administration, as well as a lower return rate of completed BSC (i.e. 86 for Total BSC at winter compared to 96 Total BSC at fall). Sensitivity probabilities were selected by identifying the highest sensitivity rating produced by each ROC curve analyses. Specificity is "the proportion of cases for which a diagnosis of disorder is rejected when rejection is warranted" (AERA, APA, & NCME, 1999, p. 182). For this study, specificity refers to the proportion of students in the sample who were identified as not at-risk on the SDQ and not at-risk on the BSC. Specificity of the BSC ranged from .58 (Winter BSC) to .67

(Fall and Spring BSC). These probabilities suggest 58% to 67% of students identified as not at-risk on the SDQ at the end of the school year, were also found not at-risk on the BSC throughout the course of the school year. Again, these specificity probabilities were selected by identifying the highest sensitivity and maximizing the specificity rating produced by the ROC curve analyses.

The other option for reporting the sensitivity and specificity of the BSC, was to use the cut score proposed by the authors of the BSC in the original study, which was 36 (Muyskens et al., 2007). In addition, King and Reschly (2014) used a cut score of 27 to adjust using only 11 of the 12 BSC items and to identify the top 20% of students who may be at-risk. If either one of these cut scores were selected and used in this study, it would have drastically impacted the sensitivity and specificity. For example, if a cut score of 36 was used for the Fall BSC administration, this would have generated significantly worse sensitivity (.29) while simultaneously improving specificity (.98) probabilities. Using the proposed cut score in the original study would have negatively impacted the number of students identified as at-risk for experiencing behavioral difficulties. If the cut score of 27 was used, this would have produced poor sensitivity (.61) and excellent specificity (.95) probabilities. In both of these cases, the specificity would be excellent, but at the expense of producing poor sensitivity probabilities. These examples highlight the importance and value of conducting classification accuracy analyses for universal screening measures.

Positive and Negative Predictive Values

The last two quality indicators to examine is the PPV and NPV. Both are important to consider because they are influenced by the prevalence of the disorder measured, or in this study, the prevalence of risk-status. Lambert et al. (2014) suggest PPV and NPV offer "a more contextualized perspective of the diagnostic quality" or screening students at-risk with the BSC (p. 56). For the current study, PPV is the proportion of students at-risk on the BSC and who are at-risk on the SDQ. Although PPV appears similar to sensitivity, it is calculated using students at-risk on the BSC and on the SDQ. Sensitivity was calculated using students identified at-risk on the SDQ regardless of their risk status on the BSC. The PPV for the BSC was low, with probabilities ranging from 47% (Winter BSC) to 57% (Fall BSC).

The results for PPV are different than sensitivity probabilities and suggest that students classified as at-risk on the BSC were classified not at-risk on the SDQ, resulting in a higher rate of false positives. When determining cut scores by maximizing sensitivity and specificity, and investigating the prevalence rate (positive and negative predictive power), it is important to consider the context in which the universal screening measure is being used for and what is the overarching goal of implementing a screener. In some areas, such as academic screening, a high rate of false positives may seem less of an issue when compared to identifying children with early signs of a disruptive behavior disorder. Strong consideration should be given to understanding the plausible impact of high false positives compared to high false negatives. A question that should drive the decision making process of determining cut scores: What are the

potential consequences when incorrectly labeling a student at-risk compared to the potential consequences of incorrectly identified a student not at-risk? This is a difficult question to answer that does not present with a clear answer. Furthermore, the answer will likely change depending on the context and the outcome of a screening measure. In the schools when looking at behavior screeners, over-identification (or false positives) will result in more students receiving Tier 2 supports in the classroom when in fact, these services may not actually be warranted. For schools who have limited resources, this is problematic. High rates of false positives also present a unique challenge because there is a potential of being identified at-risk and being incorrectly "labeled". There continues to be a stigma surrounding mental health, whether it is an internalizing disorder (i.e. anxiety) or an externalizing disorder (i.e. oppositional defiant disorder). Therefore, within the context of schools and universal behavior screening, high false positives may not necessarily be ideal.

Negative predictive value is the proportion of students not at-risk on the BSC who were also identified not at-risk on the SDQ. This appears similar to specificity, but the calculations are very different just as with sensitivity and PPV. Negative predictive values are calculated using students not at-risk on the BSC and not at-risk on the SDQ, whereas specificity was calculated using students identified not at-risk only on the SDQ regardless of their risk status on the BSC. The NPV for the BSC was much higher than the PPV, with probabilities ranging from 90% (Winter BSC) to 98% (Spring BSC). The NPV probabilities reported in this study suggest students classified as not at-risk on the BSC are also classified as not at-risk on the SDQ (true negatives). Based on the NPV

results, the BSC performs better at discriminating between students who are not at-risk compared to discriminating from students who are at-risk, when accounting for risk status on both screening measures.

When understanding the classification accuracy of the BSC, there are some notable strengths and weaknesses. At first glance, the BSC does an excellent job of discriminating students at-risk compared to those not at-risk, based on the AUC and sensitivity probabilities. This is helpful for schools who want to use the BSC as a universal screening measure to identify students who may need additional support, above and beyond what is provided at the universal level. Given the excellent sensitivity, the BSC can accurately identify students at-risk for experiencing behavioral challenges in the classroom. However, when accounting for the risk status on both screening measures, students were over-identified at-risk according to the BSC. This information is valuable because schools often do not have an abundance of resources available to provide additional support or intervention services to students when unwarranted. There are also potential negative consequences that should be considered when students are incorrectly labeled at-risk compared to being incorrectly labeled not at-risk for future disruptive behavior disorders. The BSC may be used as an initial screening tool or part of a multiple gated screening system to flag students who may be at-risk. Follow up measures could be administered such as classroom observations, teacher interview, or monitoring of academic performance to further distinguish those who need support from those who do not need support.

Research Question #6

Do the BSC Total scores demonstrate stronger classification accuracy in identifying students at-risk compared to ODRs and student absences? To answer this research question, additional ROC curve analyses were conducted with the predictor variables that were found to be statistically significant in the multiple regression analyses. Based on the results, the BSC demonstrates stronger classification accuracy compared to Winter ODRs. In looking at the AUC, Winter ODRs was not statistically significant and is able to predict at-risk behavior problems no better than chance (AUC = .59). These findings extend the current literature and are similar to the classification accuracy results published by Miller et al. (2015), where ODRs across time failed to yield statistical significance. Given the similarity of findings across both studies, schools should consider using ODRs more of a guide to understanding system level concerns rather than predicting individual student outcomes.

As many schools often use ODRs to guide identification of students who may be at-risk for experiencing emotional and behavior problems in the classroom, this may not be an appropriate source of data. ODRs were initially developed to monitor the overall trend of behavior concerns within a school system, not necessarily to predict individual outcomes for students. The lack of statistically significant findings may be the result of age and developmental expectations, as kindergarten students are often not referred to the office as frequently when compared to older students. Based on the current findings, it is recommended to not use ODRs for predicting individual student outcomes. However, additional research should continue to focus on understanding the technical

adequacy (e.g. classification accuracy) of ODRs with a larger sample of kindergarten students and examine how ODRs function across grade levels or age.

Limitations

As with any research study, there are limitations that warrant discussion. First, the sample size for this study was relatively small. Although there were several significant findings in many of the research questions, variables such as ODRs (except Winter) and number of absences were not statistically significant predictor variables. This contrasts the King and Reschly (2007) study, where both variables were statistically significant predictors of the criterion measure used. However, they had over 22,000 students participate in their study. Sample size (along with sampling error) greatly influences the ability to achieve statistical significance (Thompson, 1994). Small sample size may be one explanation as to why ODRs and/or student absences were not statistically significant predictor variables for this study.

In addition to a small sample size, the return rate for parent consent form was low. With approximately 275 potential students (11 classrooms, with 25 students in each classroom), only 96 returned consent forms. Subsequently, there was variability of returned consent rates across classrooms. For example, one classroom had three students return a signed consent form (out of 25), whereas another classroom had 15. This is a limitation to the findings of this study because it is unknown why some parents returned a consent form and why other parents did not. Often times in research studies, those who would most benefit from participation opt out. In this study, it is possible that students who presented with the greatest at-risk behaviors in the classroom may have not

been captured because their parents failed to sign a consent form. Stigma of being identified or "labeled" is also always a concern as well, when exploring behavior screeners in the classroom. Some parents may be hesitant to allow their child to be included in research on behavior screening measures for the mere fact that their child could be identified as "at-risk". Furthermore, some of the teacher participants were more involved with the consent process, as they sent home reminders to parents. The lack of return consent forms may also have been due to the level of involvement of the classroom teacher. Regardless of the reason, the low return rate for parent consent forms is a limitation to this study and it is possible that the results are biased towards those students who did return the parent consent form.

The sample also included a portion of kindergarten students from a larger pool of students. It would be ideal to include an entire grade level, not just a select sample of students from a grade level. The student participants were enrolled in classes that were hand selected by school district administrators. There were approximately 800 students in kindergarten at this school district; however, school officials only selected three of their campuses to be included in the study. The remaining campuses were excluded for reasons unknown. There may be some type of selection bias or participation bias the researcher is unaware of. Furthermore, not all students in each of the selected classrooms were included in the study. They were only included if their parents signed the consent form. As each of the classrooms had approximately 25 students enrolled, some had only a few (e.g. 3) return the consent form. There is a possibility that students who were not included in the study, were experiencing more significant emotional and

behavioral concerns. There is no way to determine this, unless all students in a classroom, school, or district are included in the screening process.

The location of the school district and demographics of the participants is another limitation. This study was conducted with a medium sized school district, and only three elementary campuses of the district were included. This school district resides in an industrial region in the south, which may not be representative of schools across the country. The sample also included only kindergarten students. Given the demographics of the sample, the results may not generalize to other districts that are perhaps more urban or rural. This was the first study to explore classification accuracy with the BSC, so it is difficult to generalize the findings to other schools or even other demographics of students.

Risk status for individual students was essentially determined using one screening measures, completed by teachers. As teacher ratings were the focus of this study, this is a limitation that should be addressed in future studies. As with any type of behavior rating scale, teacher ratings may have had some biases in their responses. Depending on the individual student and the quality of relationship the teacher has with the student, there is always a potential for teacher ratings to reflect their personal biases towards each student. If a teacher views a student more favorably, or if they know the student's family outside of school, they may rate the child's behaviors more positively than they should. However, teacher biases also may reflect more negatively, if the teacher does not necessarily like a particular student. Negative or skewed biases are often problematic when attempting to identifying students who exhibit challenging

behaviors because these are the students that are the most disruptive in the classroom. There is really no way to capture potential biases or examine how prevalent it was within the data collected for this study. The BSC does not have validity scales built within the items to monitor various response patterns. Therefore, it is recommended to consider the potential of teacher biases when interpreting the results and how they may be impacted. ODRs are also considered discretionary placements and the type of offenses or reasons why students were sent to the office was not provided. There is controversy when using ODRs with elementary students (particularly, kindergarten students) because of the lack of evidence to support validity. For example, Nelson et al. (2002) found that only 3% of elementary students who received ORDs also presented with aggressive behaviors that were clinically significant, compared to 42% of middle and high school students. ODRs may not be a valid representation of behavioral functioning in the classroom when it comes to kindergarten students.

Future Research

As this study extends the current literature regarding the validity of the BSC, as well as preliminary evidence of classification accuracy, there are several avenues for future research to explore. It is recommended that future studies of the BSC include a more realistic number of students in the data analysis and screen all students in a grade level or school. It would be valuable to include other grades/ages of student participants or students with a range of special education disabilities. Other considerations to include would be to conduct analyses that differentiate students by their ethnicity or gender. By

addressing these recommendations in future studies, this would increase generalization of results.

Future studies may also consider the inclusion of other sources of data to further demonstrate technical adequacy of the BSC. This may include structured classroom behavior observations, parent rating scales, parent interviews, teacher interviews, student interviews, or even academic performance on standardized assessments or district bench marking. By including academic achievement, one could investigate the degree to which students who are identified at-risk on the BSC also experience academic impairment. This type of information is extremely valuable for school districts and may even facilitate further buy-in for schools to begin utilizing class wide screening measures (depending on the results).

Despite the extensive amount of research available to support the psychometric properties of the SDQ, this is considered a universal screening measure. It may be of interest to include a criterion measure that is diagnostic or more comprehensive. This may include the Behavior Assessment System for Children (3rd Ed.; BASC-3) developed by Reynolds and Kamphaus (2015). By including a measure such as the BASC-3, additional data could be analyzed to investigate the psychometric properties of the BSC (e.g. classification accuracy). As this study was the first to analyze BSC data throughout the duration of one academic year, future studies may consider exploring how well the BSC functions over the course of multiple years. In doing so, one would hopefully be able to further validate the technical adequacy of the BSC. It is recommended to begin with a sample of participants starting earlier in their academic career (i.e. preschool) and

follow them throughout the elementary school years. This could assist in determining if different cut scores are appropriate for different grade levels and ages.

Practical Implications

As schools continue to implement universal screening practices, it is vital that available screening measures are able to accurately distinguish between students at-risk from those who are not at-risk. More recently, research has focused on conducting advanced statistical analyses such as a ROC curve, rather than relying solely on correlation coefficients or multiple regression analyses. Not to diminish the importance of these types of analyses, but results from a ROC curve offers valuable information where other analyses are limited. In using ROC curve analysis to determine the classification accuracy of a universal screening measure, schools may be able to adjust their cut scores to fit their student population. As previously discussed, using the recommended cut score proposed by the authors of the BSC would have negatively impacted the sensitivity and specificity. In doing so, several students would not have been identified as at-risk and the likelihood of them receiving Tier 2 supports would substantially decrease.

This study highlights the importance of understanding all of the classification accuracy indicators. At first glance, the AUCs and sensitivity probabilities provide evidence that the BSC is an excellent screening measure. Furthermore, the BSC has high sensitivity and high NPV. Based on a combination of these indicators, it appears that the BSC is able to accurately differentiate between students at-risk from those who are not at-risk, with a low rate of false negatives. However, the low specificity

probabilities and low PPV suggest the BSC has higher rates of false positives. When considering all the different classification accuracy indicators, it is better for a screening measure to have high sensitivity and low specificity, than having low sensitivity and high specificity. Given the overall result, this study offers preliminary evidence of the BSC's classification accuracy. As more research is conducted, researchers and practitioners will be able to better understand the psychometric properties and utility of the BSC.

Conclusions

The purpose of this study was to explore the BSC as a universal screening measure with a sample of kindergarten students. Conventional analyses (e.g. bivariate correlation, multiple regression) were conducted to explore the consistency and validity of the BSC. In addition, several ROC curve analyses were conducted in order to explore the classification accuracy of the BSC as a universal screening measure. This study was the first to utilize classification accuracy indicators to investigate the degree to which the BSC is able to accurately differentiate between students at-risk from students not at-risk for experiencing behavior difficulties in the classroom. Current research on the BSC suggests it is a reliable and valid screening measure to use in the schools; however, the evidence is limited to correlation coefficients and multiple regression coefficients.

Results from this study conclude that the BSC yields consistent teacher ratings over time. Convergent and predictive validity were also supported, using the SDQ as a criterion measure. Multiple regression analyses revealed BSC scores accounted for a statistically significant amount of variance in year-end SDQ scores. This was a finding

that occurred across all three multiple regression models. Winter ORDs was also a statistically significant predictor variable of Spring SDQ. Overall, the BSC is able to differentiate between those at-risk from those not at-risk. Classification accuracy indicators, suggest the BSC exhibits good to excellent sensitivity but poor specificity. Of the predictor variables that were statistically significant, Spring BSC scores resulted in the strongest classification accuracy based on the AUC statistics. As this was the first study to explore the classification accuracy of the BSC, it provides a foundation for understanding how well the BSC functions as a universal screening measure when used in a classroom setting.

REFERENCES

Achenbach, T. M., & Edelbrock, C. (1987). *Manual for the Teacher Report Form of the Child Behavior Checklist*. Burlington: Queen City Printers.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms and Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Albers, C. A., & Kettler, R. J. (2014). Best practices in universal screening. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp. 121-131). Bethesda, MD: NASP Publications.

Alexander, K. L., Entwisle, D. R., Horsey, C. S. (1997). From first grade forward: Early foundations of high school dropouts. *Sociology of Education, 70*, 87-107.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (1999). *The standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Bayat, M., Mindes, G., & Covitt, S. (2010). What does RTI (Response to Intervention) look like in preschool? *Early Children Education Journal, 37*, 493-500. doi: 10.1007/s10643-010-0372-6

Behar, L., & Stringfield, S. (1974). *Manual for the Preschool Behavior Questionnaire.* Durham, NC: Behar.

Bezdek, J. M. (2014). *An examination of the validity of office disciplinary referrals*

(Doctoral dissertation). Retrieved from Psychinfo. (UMI: 3454497).

Bradley, R., Danielson, L., & Doolittle, J. (2007). Responsiveness to intervention 1997 to 2007. *Teaching Exceptional Children, 39*(5), 8-12.

Bradley, R., Doolittle, J., & Bartolotta, R. (2008). Building on the data and adding to the discussion: The experiences and outcomes of students with emotional disturbance. *Journal of Behavioral Education, 17*, 4-23.

Bradshaw, C. P., Koth, C. W., Thorton, L. A., & Leaf, P. J. (2009). Altering school climate through school-wide positive behavioral interventions and supports: Findings from a group-randomized effectiveness trial. *Prevention Science, 10*, 100-115.

Campbell, S. B. (1995). Behavior problems in preschool children: A review of recent research. Journal of Child Psychology and Psychiatry, 36(1), 113-149. doi: 10.1111/jcpp.1995.36.issue-1

Carroll, H. M. C. (2013). The social, emotional and behavioral difficulties of primary school children with poor absences records. *Educational Studies, 39,* 223-234. doi: 10.1080/03055698.2012.717508

Catalano, R. F., Haggerty, K. P., Osterle, S., Fleming, C. B., & Hawkins, J. D. (2004). The importance of bonding to school for healthy development: Findings from the social development research group. *Journal of School Health, 74*, 252-261.

Chafouleas, S. M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education & Treatment of Children, 34*, 575–591. doi:10.1353/etc.2011.0034

Chafouleas, S. M., Kilgus, S. P., & Wallach, N. (2010). Ethical dilemmas in school-based behavioral screening. *Assessment for Effective Intervention, 35*, 245-252. doi: 10.1177/1534508410379002

Commission on Chronic Illness (1957). *Chronic illness in the United States: Volume I. Prevention of chronic illness.* Cambridge, MA: Harvard University Press.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in the first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*, 394 - 409.

Conners, C. K. (1989). *Manual for the Conners' Rating Scales.* North Tonawanda, NY: Multi-Health Systems.

Cook, C. R., Rasetshwane, K. B., Truelson, E., Grant, S., Dart, E. H., Collins, T. A., & Sprague, J. (2011). Development and validation of the Student Internalizing Behavior Screener: Examination of reliability, validity, and classification accuracy. *Assessment for Effective Instruction, 36*(2), 71-79. doi: 10.1177/1534508410390486

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cullinan, D., & Epstein, M. H. (2013). Development, reliability, and construct validity of the Emotional and Behavioral Screener. *Preventing School Failure: Alternative Education for Children and Youth, 57*, 223-230. doi: 10.1080/1045988X.2012.715356

Dever, B. V., Dowdy, E., Raines, T. C., & Carnazzo, K. (2015). Stability and change of behavioral and emotional screening score. *Psychology in the Schools, 00*, 1 -11. doi: 10.1002/pits.21825

Dever, B. V., Rains, T. C., & Barclay, C. M. (2012). Chasing the unicorn: Practical implementation of universal screening for behavioral and emotional risk. *School Psychology Forum: Research in Practice, 6*, 108-118.

DiStefano, C. A., & Kamphaus, R. W. (2007). Development and validation of a behavioral screener for preschool-age children. *Journal of Emotional and Behavioral Disorders, 15*(2), 93-102.

Dowdy, E., Dever, B. V., Raines, T. C., & Moffa, K. (2016). A preliminary investigation into the added value of multiple gates and informants in universal screening for behavioral and emotional risk. *Journal of Applied School Psychology, 32*(2), 178-198, doi: 10.1080/15377903.2016.1165327

Dowdy, E., Furlong, M., Eklund, K., Saeki, E., & Ritchey, K. (2010). Screening for mental health and wellness: Current school based practice and emerging possibilities. In B. Doll, W. Pfohl, & J. Yoon. (Eds.), *Handbook of youth prevention science* (pp. 70-95). New York, NY: Routledge.

Dowdy, E., Nylund-Gibson, K., Felix, E. D., Morovati, D., Carnazzo, K. W., & Dever, B. V. (2014). Long-term stability of screening for behavioral and emotional risk. *Educational Psychological Measurement, 74*, 453-472. doi: 10.1177/0013164413513460

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

Eber, L., & Nelson, C. M. (1997). School-based wraparound planning: Integrating services for students with emotional and behavioral needs. *American Journal of Orthopsychiatry, 67*, 385–395.

Eklund, K., Renshaw, T. L., Dowdy, E., Jimerson, S. R., Hart, S. R., Jones, C. N., & Earhart, J. (2009). Early identification of behavioral and emotional problems in youth: Universal screening versus teacher-referral identification. *The California School Psychologist, 1494*, 89-95.

Elliot, S. N., & Gresham, F. M. (2007). *Social Skills Improvement System: Performance Screening Guides*. Bloomington, MN: Pearson Assessments.

Ennis, R. P., Lane, K. L., & Oakes, W. P. (2012). Score reliability and validity of the student risk screening scale: A psychometrically sound, feasible tool for use in urban elementary schools. *Journal of Emotional and Behavioral Disorders, 20*, 241-259. doi: 10.1177/1063426611400082

Epstein, J. L., & Sheldon, S. B. (2002). Present and accounted for: Improving student attendance through family and community involvement. *The Journal of Educational Research, 95*, 308-318. doi: 10.1080/00220670209596604

Epstein, M. H. (2004). *Behavioral and Emotional Rating Scale: A strengths-based approach to assessment (2nd Edition).* Austin, TX: PRO-ED.

Epstein, M. H., & Cullinan, D. (2010). Scales for Assessing Emotional Disturbance (2nd ed.). Austin, TX: PRO-ED.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27,* 861-874. doi: 10.1016/j.patrec.2005.10.010

Feil, E. G., Severson, H. H., & Walker, H. M. (1995). Identification of critical factors in the assessment of preschool behavior problems. *Education and Treatment of Children, 18,* 261-271.

Feil, E. G., Severson, H. H., & Walker, H. M. (1998). Screening for emotion and behavior delays: The Early Screening Project. *Journal of Early Intervention, 21,* 252-266.

Feil, E. G., Walker, H. M., Severson, H. H., & Ball, A. (2000). Proactive screening for emotional/behavioral concerns in Head Start preschools: Promising practices and challenges in applied research. *Behavioral Disorders, 26,* 13-25.

Freeman, J., Simonsen, B., McCoach, B., Sugai, G., Lombardi, A., & Horner, R. (2016). Relationship between school-wide positive behavior interventions and supports and academic, attendance, and behavior outcomes in high schools. *Journal of Positive Behavior Interventions, 18(1),* 41-51. doi: 10.1177/1098300715580992

Fuchs, L. S., & Fuchs, D. (2007). A model implementing responsiveness to intervention. *Teaching Exceptional Children, 39*(5), 14-20.

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

Gresham, F. M., & Elliot, S. N. (1990). *The Social Skills Rating System.* Circle Pines, MN: American Guidance Service.

Gresham, F. M. (2008). Best practices in diagnosis in a multitier problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (pp. 281-294). Bethesda, MD: National Association of School Psychologists.

Gresham, F. M., & Elliot, S. N. (2008). *Social Skills Improvement System Rating Scales.* San Antonio, TX: Pearson.

Gresham, F. M., Hunter, K. K., Corwin, E. P., & Fischer, A. J. (2013). Screening, assessment, treatment, and outcome evaluation of behavioral difficulties in an RTI model. *Exceptionality: A Special Education Journal, 21*, 19-33. doi: 10.1080/09362835.2013.750115

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

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry, 177*, 534-539.

Gottfried, M. A. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis, 31*, 392–415.

Harber, J. R. (1981). Assessing the quality of decision making in special education. *The Journal of Special Education, 15*, 77-90.

Harris, R., Sawaya, G. F., Moyer, V. A., & Calonge, N. (2011). Reconsidering the criteria for evaluating proposed screening programs: Reflections from four current and former members of the U.S. Preventive Services Task Force.

Epidemiological Reviews, 33, 20-35. doi: 10.1093/epirev/mxr005

Hawken, L. S., Vincent, C. G., & Shumann, J. (2008). Response to intervention for social behavior: Challenges and opportunities. *Journal of Emotional and Behavioral Disorder, 16*, 213-225.

Hill, C. R., & Hughes, J. N. (2007). An examination of the convergent and discriminate validity of the Strengths and Difficulties Questionnaire. *School Psychology Quarterly, 22*, 380-406.

Jaffee, S. R., Harrington, H., Cohen, P., & Moffitt, T. E. (2005). Cumulative prevalence of psychiatric disorders in youths. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*, 406-407.

Kamphaus, R. W., DiStefano, C., Dowdy, E., Eklund, K., & Dunn, A. R. (2010). Determining the presences of a problem: Comparing two approaches for detecting youth behavioral risk. *School Psychology Review, 39*, 395-407.

Kamphaus, R. W., & Reynolds, C. R. (2007). *BASC-2: Behavioral and Emotional Screening System (BESS).* Minneapolis, MN: Pearson Assessments.

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

Kearney, C. A. (2008). School absenteeism and school refusal behavior in youth: A contemporary review. *Clinical Psychology Review, 28*, 451- 471. doi:10.1016/j.cpr.2007.07.012

Kelm, J. L., & McIntosh, K. (2012). Effects of school-wide positive behavior support on teacher self-efficacy. *Psychology in the Schools, 49,* 137-147.

King, K. R., Reschly, A. L., & Appleton, J. J. (2012). An examination of the validity of the Behavioral and Emotional Screening System in a rural elementary school validity of the BESS. *Journal of Psychoeducational Assessment, 30*, 527-538.

King, K. R., & Reschly, A. L. (2014). A comparison of screening instruments: Predictive validity of the BESS and the BSC. *Journal of Psychoeducational Assessment, 32*, 687-698. doi: 10.1177/0734282914531714

Lambert, M. C., Epstein, M. H., & Cullinan, D. (2014). The diagnostic quality of the Emotional and Behavioral Screener. *Journal of Psychoeducational Assessment, 32*(1), 51-61. doi: 10.1177/07342829134895541

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J. H., & Weisenback, J. L. (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., Jolivette, K., Conroy, M., Nelson, C. M., & Benner, G. J., (2011). Future research directions for the field of E/BD: Standing on the shoulders of giants. *Education and Treatment of Children, 34*, 423-443.

Lane, K. L., Kalberg, J. R., Lambert, W., Crnobori, M., & Bruhn, A. (2010). A comparison of systematic screening tools for emotional and behavioral disorders: A replication. *Journal of Emotional and Behavioral Disorders, 18*, 100-112.

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

Martson, D. (2002). A functional and intervention-based approach to establishing discrepancy for students with learning disabilities. In R. Bradley, D. P. Hallahan, & L. Danielson (Eds.), *Identification of learning disabilities: Research to practice*. NJ: Lawrence Erlbaum Associates, Inc.

McIntosh, K. & Goodman, S. (2016). *Integrated multi-tiered systems of support: Blending RTI and PBIS*. New York, NY: The Guilford Press.

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*, 82-91.

Miller, F. G., Cohen, D., Chafouleas, S. M., Riley-Tillman, T. C., Welsh, M. E., & Fabiano, G. A. (2015). A comparison of measures to screen for social, emotional, and behavioral risk. *School Psychology Quarterly, 30*, 184-196. doi: 10.1037/spq0000085 10

Morabia, A., & Zhang, F. F. (2004). History of medical screening: From concepts to action. *Postgrad Medical Journal, 80*, 463-469.

Morgan-D'Atrio, C., Northrup, J., LaFleur, L., & Spera, S. (1996). Toward prescriptive alternatives to suspensions: A preliminary evaluation. *Behavioral Disorders, 21*, 190 – 200.

Muyskens, P., Martson, D., & Reschly, A. L. (2007). The use of response to intervention practices for behavior: An examination of the validity of a screening instrument. *The California School Psychologist, 12*, 31-45.

Nordness, P. D., Epstein, M. H., Cullinan, D., & Pierce, C. D. (2014). Emotional and behavioral screener: Test-retest reliability, inter-rater reliability, and convergent validity. *Remedial and Special Education, 35*, 211-217. doi: 10.1177/0741932513497596

Park, S. H., Goo, J. M., & Jo, C. H. (2004). Receiver operating characteristic (ROC) curve: Practical review for radiologists. *Korean Journal of Radiology, 5(*1), 11-18.

Pas, E. T., Bradshaw, C. P., & Mitchell, M. M. (2011). Examining the validity of office discipline referrals as an indicator of student behavior problems. *Psychology in the Schools, 48*, 541-555. doi: 10.1002/pits

Pintea, S., & Moldovan, R. (2009). The receiver-operating characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychotherapies, 9*, 49-66.

Reinke, W. M., Herman, K. C., Petras, H., & Ialongo, N. S. (2008). Empirically derived subtypes of children academic and behavior problems: Co-occurrence and distal outcomes. *Journal of Abnormal Child Psychology, 36*, 759-770.

Reynolds, C. R., & Kamphaus, R. W. (2007). *Behavior Assessment System for Children, Second Edition*. Circle Pines, MN: American Guidance Service.

Reynolds, C. R., & Kamphaus, R. W. (2015). *Behavior Assessment System for Children,*

Third Edition. Circle Pines, MN: American Guidance Service.

Rimm-Kaufman, S. E., Piata, R. C., & Cox, M. J. (2000). Teachers' judgements of problems in the transition to kindergarten. *Early Children Research Quarterly*, 15(2), 147-166. doi: 10.1016/S0885-2006(00)00049-1.

Romer, D., & McIntosh, M. (2005). The roles and perspectives of school mental health professionals in promoting adolescent mental health. In D. L. Evans, E. B. Foa, R. E. Gur, H. Hendin, C. P. O'Briend, M. E. P. Seligan, & B. T. Walsh (Eds.), *Treating and preventing adolescent mental health disorders: What we know and what we don't know* (pp. 598 – 615). New York, NY: Oxford University Press.

Rumberger, R. W. (2011). *Dropping out: Why students drop out of high school and what can be done about it*. Cambridge, MA: Harvard University Press.

Ryan, J. B., Reid, R., & Epstein, M. H. (2004). Peer-mediated intervention studies on academic achievement for students with ED: A review. *Remedial and Special Education, 25*, 330–341.

Severson, H. H., Walker, H. M., Hope-Doolittle, J., Kratochwill, T. R., & Gresham, F. M. (2007). Proactive, early screening to detect behaviorally at-risk students: Issues, approaches, emerging innovations, and professional practices. *Journal of School Psychology, 45*, 193-223. doi:10.1016/j.jsp.2006.11.003

Stoep, A. V., McCauley, E., Thopson, K. A., Herting, J. R., Kuo, E. S., Stewart, D. G. & Kushen, S. (2005). Universal emotional health screening at the middle school transition. *Journal of Emotional and Behavioral Disorders, 13*, 213-223. doi: 10.1177/10634266050130040301

Stone, L. L., Otten, R., Engels, R. C. M. E., Vermulst, A. A., & Janssens, J. M. A. M. (2010). Psychometric properties of the parent and teacher version of the Strengths and Difficulties Questionnaire for 4- to 12- year olds: A review. *Clinical Child Family Psychological Review, 13,* 254-274. doi: 10.1007/s10567-010-0071-2

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, 2,* 94-102. doi:10.1177/106342660000800205

Tanner-Smith, E., E., & Wilson, S. J. (2013). A meta-analysis of the effects of dropout prevention programs on school absenteeism. *Prevention Science, 14*, 468-478.

Tilly, W. D. (2008). The evaluation of school psychology to science-based practice: Problem solving and the three-tiered model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 17-35). Bethesda, MD: National Association of School Psychologists.

Thompson, B. (1994). The concept of statistical significance testing. *Practical Assessment, Research, & Evaluation, 4*(5), 1-3.

Tobin, T. J. & Sugai, G. M. (1999). Using sixth-grade school records to predict school violence, chronic discipline problems, and high school outcomes. *Journal of Emotional and Behavioral Disorders, 7,* 40-53.

U. S. Department of Health and Human Services (1999). *Mental Health: A report of the surgeon general.* Rockville, MD: U. S. Department of Health and Human

Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institute of Health, National Institute of Mental Health.

VanDerHeyden, A. M. (2011). Technical adequacy of RTI decisions. *Exceptional Children, 77,* 1-16.

Waasdorp, T. E., Bradshaw, C. P., & Leaf, P. J. (2012). The impact of school wide positive behavioral intervention and supports on bullying and peer rejection. *Archives of Pediatrics and Adolescent Medicine, 166,* 149-156.

Walker, H. M., & Severson, H. H. (1990). *Systematic Screening for Behavior Disorders (SSBD).* Longmont, CO: Sopris West.

Walker, H. M., Small, J. W., Severson, H. H., Seeley, J. R., & Feil, E. G. (2014). Multiple-gating approaches in universal screening within school and community settings. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools,* (pp. 47-75). Washington, DC: American Psychological Association.

Wilson, J. M. G., & Junger, G. (1968). Principles and practice of screening for disease. *Public Health Papers, 34.* Retrieved from http://apps.who.int/iris/bitstream/10665/37650/1/WHO_PHP_3

Wood, J. J., Lynne-Landsman, S. D., Langer, D. A., Wood, P. A., Clark, S. L., Eddy, J. M., & Ialongo, N. (2012). School attendance problems and youth

psychopathology: Structural cross-lagged regression models in three longitudinal data sets. *Child Development, 83,* 351-366. doi: 10.1111/j.1467-8624.2011.0