Screening for Social-Emotional and Behavioral Aspects of Kindergarten Readiness: A Systematic Review of Screeners and Validation of BASC-3 BESS Teacher for Somali Students

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Alaa K. Houri

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Faith G. Miller

(June 2020)

© Alaa Houri 2020

Dedication

To my amazing daughter, Rayhan Ali, I dedicate this work to you. I hope that you one day view this success as your success, and proof that, with the support of your family, you can achieve anything you set your mind to.

Acknowledgements

I would first like to acknowledge my advisor, Dr. Faith Miller, for all of her time and effort spent guiding me throughout my last few years in graduate school. Your guidance, help, support, and encouragement helped me develop my interest and knowledge in this area of research. I truly appreciate the time and commitment you spent meeting with me and reading countless drafts of many projects. To Dr. Amanda Sullivan, thank you for believing in me and helping me develop my confidence as a future scholar. As I reflect on the past six years in this program, I realize that your ongoing support from my first day of graduate school has been instrumental in my growth as a student, and has helped me be successful in this program. To my committee member and fieldwork supervisor, Dr. Annie Hanson-Burke, thank you for always providing a safe and encouraging space for students. Your kind words of support and encouragement helped increase my confidence any time I doubted my abilities as a future school psychologist, and I hope to be able to provide that level of support to students one day. Last but certainly not least, Dr. Bonnie Klimes-Dougan, thank you for seeing my potential for success even when, as an undergraduate student, I insisted that research and academia were of no interest to me. Without your gentle push towards graduate education, and your constant belief in me, I would not be here today.

And to my family. Words cannot express the amount of gratitude and love that I have for every single one of you. To my parents, thank you for always believing that I can succeed and achieve anything that I set my mind to. To my sisters, thank you for always encouraging me and pointing out my positive traits anytime I was down, even

when doing so at odd times. To my in-laws, thank you for always going above and beyond every single time I needed your kindness and support, and for being the amazing and loving family members that you are. And to my husband. Mamdouh, you have been my rock every step of the way. You pushed me to accomplish my dreams even when I was ready to give up. Without you, I truly believe that I would not be here today. From the bottom of my heart, thank you.

The study entitled "A Systematic Review of Universal Screeners Used to Evaluate Social-Emotional and Behavioral Aspects of Kindergarten Readiness" has been published by Taylor & Francis in Journal of *Early Education and Development* on 10/30/2019, available online: https://doi.org/10.1080/10409289.2019.1677132.

Abstract

Best practices in universal screening procedures entail the use of teacher-reported screening measures for the identification of students who may benefit from social-emotional and behavioral (SEB) supports. Few studies have systematically identified available screening measures, or examined their use with diverse populations. Therefore, this project aimed to (a) examine the current landscape of teacher-reported universal screening measures for assessing the SEB components of kindergarten readiness, and (b) fill a gap in the literature on bias evaluation evidence of SEB screening measures for kindergarten students. This project, then, included two studies: Study 1 systematically identified 11 SEB screening measures and reviewed the psychometric properties of each measure. This study also included a review of bias evaluation evidence provided for each scale. Results of this study demonstrated adequate to strong reliability evidence overall; however, validity and bias evaluation evidence were severely lacking. Study 2 aimed to expand on the current research landscape for bias evaluation by examining the internal structure of the Behavioral Assessment Scale for Children, Third Edition, Behavioral and Emotional Screening Scale (BASC-3 BESS) teacher form for a Somali kindergarten student population by seeking to replicate the factor structure of the BASC-3 BESS identified for the norming population with a Somali student sample. Results of this study demonstrated similar internal structure findings between a Somali student sample and the BASC-3 BESS norming sample, providing preliminary support for its use within a Somali kindergarten student population. Future research efforts should continue

exploring the remaining psychometric properties of the BASC-3 BESS Teacher in a Somali student sample to provide further evidence for its use with this student population.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The social-emotional and behavioral (SEB) components of kindergarten readiness are frequently identified by teachers as important skills for student success, particularly because SEB functioning during kindergarten is positively predictive of future academic success and overall well-being (Abry, Latham, Bassok, & LoCasale-Crouch, 2015; Lewitt & Baker, 1995; Lin, Lawrence, & Gorrell, 2003; Montes, Lotyczewski, Halterman, & Hightower, 2011; Sabol & Pianta, 2012; West, Hausken, & Collins, 1993). Early identification of students with SEB difficulties is critical because it can facilitate SEB supports and services, leading to improvement in academic and behavioral outcomes (Albers & Kettler, 2014; Choi, Elicker, Christ, & Dobbs-Oates, 2016; Sabol & Pianta, 2012).

Best practices in universal screening strategies suggest teacher-reported screening measures are effective for early identification of students with SEB difficulties (Albers & Kettler, 2014; Ikeda, Neessen, & Witt, 2008). However, universal screening measures must demonstrate strong reliability, validity, and bias evaluation evidence, thereby indicating they are appropriate for use with all students in kindergarten classrooms. This dissertation, then, has two purposes: (1) to examine the current landscape of teacher-reported universal screening measures for the identification of students with SEB difficulties, and (2) to expand on the existing literature reporting on the bias evaluation evidence of SEB screening measures for students of diverse racial backgrounds. In order to accomplish these goals, a review of the components of SEB functioning as well as universal screening is first required.

Background

The SEB components of kindergarten readiness are of great importance for students' overall wellbeing. Specifically, students with strong SEB skill development at kindergarten entry demonstrate stronger language, motor, and academic skills over time (Montes et al., 2011; Sabol & Pianta, 2012). Conversely, students who display problem behaviors in kindergarten may be at increased risk for academic and behavioral difficulties over time (Breslau et al., 2009). Therefore, it is critical that SEB components of kindergarten readiness are clearly defined in order to accurately screen for and identify the SEB components that can be strengthened through additional supports.

SEB dimension of kindergarten readiness. The SEB components of kindergarten readiness positively linked to future well-being include self-regulation and social functioning skills (Blair & Razza, 2007; Zelazo & Müller, 2002; Zelazo, Qu, & Kesek, 2010). Specifically, these constructs include the ability to regulate emotions, express behaviors, and form relationships with others in culturally and socially appropriate ways (Alfonso & Flanagan, 2009; Collaborative for Academic, Social, and Emotional Learning [CASEL], 2018; Halle & Darling-Churchill, 2016; Kagan, Moore, & Bradekamp, 1995; National Education Goals Panel, 1997; Yates et al., 2008). For the purposes of this study, then, SEB functioning was defined to include students' emotional self-regulation, behavioral self-regulation, and social and interpersonal skills.

Taken together, these SEB constructs support and strengthen academic and SEB development. Specifically, students who are better able to regulate their emotions demonstrate stronger behavioral self-regulation skills, leading to more frequent learning behaviors such as staying on task and following directions (Edossa, Schroeders, Weinert,

& Artelt, 2018; Howse, Lange, Farran, & Boyles, 2010; McClelland & Cameron, 2011). Furthermore, emotional and behavioral self-regulation skills may positively predict prosocial encounters with others in the classroom (Edossa et al., 2018; McClelland & Cameron, 2011). Consequentially, students with stronger prosocial skills are more likely to adhere to culturally appropriate communication methods such as requesting and obtaining help, support, and encouragement from others, which may ultimately influence academic achievement (Coolahan, Fantuzzo, Mendez, & McDermott, 2000). Moreover, emotional and behavioral self-regulation as well as social and interpersonal skills work together to promote future academic achievement (Howse, Calkins, Anastopoulos, Keane, & Shelton, 2003; Li-Grining, Vortruba-Drzal, Maldonado-Carreno, & Haas, 2010; Liew, 2012). Therefore, assessment of these skills may inform efforts to set up students for success upon school entry. To this end, universal screening measures that follow best practices guidelines can effectively identify students with SEB difficulties who may benefit from additional supports.

Universal screening of SEB constructs of school readiness. Unfortunately, most schools in the United States are not implementing thorough universal screening procedures for SEB functioning (Bruhn, Woods-Groves, & Huddle, 2014). Instead, schools often rely on other methods for identifying students with SEB difficulties, such as the use of existing data (e.g., office discipline referrals, attendance) or teacher and staff reports (Bruhn et al., 2014; Miller et al., 2015; Severson et al., 2007). However, these methods are not as reliable in identifying students with SEB difficulties as psychometrically-sound rating scales, and have also shown to increase disproportionality throughout the referral process (Dever, Raines, Dowdy, & Hostutler, 2016; Dowdy,

Kamphaus, Twyford, & Dever, 2014; Eklund et al., 2009; Miller et al., 2015; Raines, Dever, Kamphaus, & Roach, 2012).

Best practices in universal screening entails screening for SEB skills through the use of screening measures alongside academic achievement screening procedures (Essex et al., 2009; Glover & Albers, 2007; Ikeda et al., 2008). Such procedures ensure all students requiring additional SEB supports are identified and may reduce the likelihood of disproportionately identifying students for services (Dever et al., 2016; Dowdy et al., 2014; Raines et al., 2012). For this reason, it is crucial that the identified universal screening measures possess strong psychometric properties, including the use of a representative norming sample and thorough bias evaluation procedures (Glover & Albers, 2007).

Rationale

Numerous studies have reviewed screening measures for SEB constructs; however, many do not include an exhaustive systematic review of available SEB rating scales, and may instead review a handful of popular rating scales (for examples, see Carter, Briggs-Gowan, & Davis, 2004; Halle & Darling-Churchill, 2016; Jenkins et al., 2014; Oakes, Lane, & Ennis, 2016; Oakes, Lane, Cox, & Messenger, 2014). Therefore, this dissertation addresses this gap by first completing a systematic review of available teacher rating scales of the identification of students with SEB difficulties.

Study 1. Given the centrality of these issues to student success, researchers have assembled and reviewed many rating scales that can be used for the purposes of screening for SEB difficulties for students of all ages (Halle & Darling-Churchill, 2016; Jenkins et al., 2014; Severson et al., 2007; Yates et al., 2008). In addition, several compendia

provide lists and reviews of available screeners for use with the kindergarten population (Brown, Scott-Little, Amwake, & Wynn, 2007; Niemeyer & Scott-Little, 2002). However, systematic reviews of scales directly related to the SEB components of kindergarten readiness, specifically, are scant. In addition, researchers may not consistently include cultural bias reviews when systematically reviewing rating scales. Therefore, study 1 systematically reviewed universal screeners used to evaluate SEB constructs of kindergarten readiness. Furthermore, a thorough review of the psychometric properties, including a review of bias identification methods, was provided for each identified rating scale. To this end, study 1 answered the following research questions:

1. What are the available universal teacher-report screeners that measure the SEB constructs of school readiness that are currently available for use in a classroom setting for kindergarten students?
2. What are the psychometric characteristics listed for each of the scales?
3. To what extent are the scales appropriate for use with diverse student populations?

The results of study 1 identified a total of 11 appropriate screening measures that demonstrated adequate to strong reliability evidence and inadequate to no evidence of validity and bias evaluation methods for most measures. The findings of this study suggested a need for more thorough evaluation of validity and bias of test scores for SEB screening in diverse students. For this reason, study 2 fills a much-needed gap by exploring various psychometric properties of the Behavior Assessment System for Children, Third Edition Behavioral and Emotional Screening System (BASC-3 BESS) teacher for a specific culturally-diverse kindergarten population. Although the BASC-3

BESS has been identified as the most frequently used SEB screening scale in the U.S. (Bruhn et al., 2014), limited psychometric properties have been provided particularly in relation to its use with diverse student populations.

Study 2. Although the BASC-3 BESS manual demonstrated adequate reliability, as demonstrated by the results of study 1, validity and bias evaluation evidence were lacking (Hogan, 2017; Sink & Carlisle, 2017). Study 2 expanded the validity and bias evaluation evidence of the BASC-3 BESS by providing validity evidence based on its internal structure specifically with a Somali kindergarten student population. Confirmatory factor analyses were conducted to determine if items map onto each of the three latent factors with the Somali student population as they did with the norming sample (Kamphaus & Reynolds, 2015). The internal structure of the BASC-3 BESS was also explored through a series of Spearman correlations that examined the interrelationships among the item and subindex scores and between the subindex scores and the total composite score. To this end, study 2 answered the following research questions:

1. Does the higher-order single-factor model of the BASC-3 BESS Teacher identified by the test developers (Kamphaus & Reynolds, 2015) result in a well-fitted model when applied to a Somali kindergarten population as determined by the goodness of fit indices and individual factor loadings?
2. Do the BASC-3 BESS Teacher items identified within each latent factor correlate with each other for the Somali student population as demonstrated by a correlation of 0.80 or greater (Alfonso & Flanagan, 2009; Bracken, 1987), indicating adequate internal consistency?

3. Do the subindex scores correlate with the BERI score with the Somali student population with correlations above 0.70 as they do with the norming population (Kamphaus & Reynolds, 2015)?

Results of study 2 demonstrated adequate reliability and validity based on the internal structure of the BASC-3 BESS Teacher with a kindergarten Somali student population, providing preliminary support for its use within a Somali kindergarten student population.

Conclusions

Due to the limitations of extant literature, this line of research is a first step in improving the field's knowledge on screening tools for SEB functioning with diverse populations. This project contributes to the field by identifying appropriate universal screening measures for individual schools' student populations and providing preliminary evidence for the use of a universal SEB screening measure in a Somali kindergarten student population. Together, these two studies inform practice through supports in identifying appropriate universal screening measures for their student population.

Definition of Terms

- **Behavioral self-regulation.** A sub-component of self-regulation, behavioral self-regulation refers to the behaviors that are related to engagement or learning, and are defined as the ability to suppress dominant behaviors and enact subdominant less-desirable responses such as following directions, taking turns during group activities, and staying on task as needed (Blair, 2002; Blair & Razza, 2007; Bodrova & Leong, 2008; Linder, Ramey, & Zambak, 2013).
- **Bias.** Any differences in results or test scores of a screening measure that is not related to students' ability but is related to students' specific demographic variables and may lead to differences in test score interpretations (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Rios & Wells, 2014).
- **Confirmatory factor analysis.** Type of analysis that measures the relationship between observed measures, such as items on a rating scale, and latent variables (Brown, 2015).
- **Latent factor.** Can also be referred to as *latent structure.* Unobserved characteristics that can be measured through observable behaviors or characteristics (Schreiber, Nora, Stage, Barlow, & King, 2006). An example of a latent factor would be internalizing behaviors.
- **Emotional self-regulation.** A sub-component of self-regulation, emotional self-regulation refers to one's ability to control their emotions (Edossa et al., 2018). Such skills include the ability to assess the environment or situation that one encounters,

and identify and subsequently enact the appropriate emotional response (Raver et al., 2010).

- **Social-emotional and behavioral functioning.** A multi-faceted construct related only to the social, emotional, and behavioral components of school readiness, and includes students' emotional self-regulation, behavioral self-regulation, and social and interpersonal skills.
- **Social and interpersonal skills.** Set of skills required for successful interaction and the formation of strong, healthy, and supportive relationships with both peers. These include communication skills, cooperation, and respect for others (CASEL, 2018; DiPerna, Volpe, & Eliott, 2002; Ray & Smith, 2010).
- **Universal screening.** The process of identifying students who are experiencing academic and SEB difficulties who may be considered at-risk and in need of additional services (Glover & Albers, 2007).
- **Validity based on internal structure.** Type of validity that provides information on the extent to which the test items correlate with the constructs of the proposed interpretation of the scale (AERA et al., 2014; Rios & Wells, 2014).

Chapter 2

Study 1: A Systematic Review of Universal Screeners Used to Evaluate Social-Emotional and Behavioral Aspects of Kindergarten Readiness

The implementation of universal screening for kindergarten readiness skills early in kindergarten is important for identifying students who may require additional supports to promote academic success (Ikeda, Neessen, & Witt, 2008). These universal screening strategies can result in early intervention supports for students with minimal kindergarten readiness skills in order to strengthen their skills, thereby increasing their likelihood for future academic and social, emotional, and behavioral success (Albers & Kettler, 2014; Quirk, Grimm, Furlong, Nylund-Gibson, & Swami, 2016; Reinke, Herman, Petras, & Ialongo, 2008; Snow, 2016). Kindergarten readiness skills include students' overall physical well-being, motor and language development, social-emotional and behavioral (SEB) development, approaches to learning skills, and general knowledge and cognitive development (Kagan, Moore, & Bradekamp, 1995; National Education Goals Panel [NEGP], 1997; Snow, 2006).

The SEB dimension of kindergarten readiness is frequently identified as important to kindergarten teachers above and beyond other dimensions of kindergarten readiness (Abry, Latham, Bassok, & LoCasale-Crouch, 2015; Lewitt & Baker, 1995; Lin, Lawrence, & Gorrell, 2003; West, Hausken, & Collins, 1993). Strong SEB skill development at kindergarten entry is positively predictive of future academic performance (Sabol & Pianta, 2012), whereas poor SEB functioning is negatively related to future achievement (Breslau et al., 2009). Fortunately, research suggests early intervention for SEB functioning is positively associated with stronger kindergarten

academic entry skills and greater academic growth over time (Choi, Elicker, Christ, & Dobbs-Oates, 2016). Thus, the focus on universal screening for the SEB dimension of kindergarten readiness is warranted.

There are many available screeners for the identification of SEB functioning in schools (Halle & Darling-Churchill, 2016; Jenkins et al., 2014; Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007; Yates et al., 2008). However, school administrators and staff often lack the time and finances required to review all available screeners and identify the optimal one in terms of its intended purpose and psychometric properties. This challenge is compounded by the fact that many available screeners are not appropriate for the identification of SEB concerns at kindergarten entry or may not be appropriate for diverse student populations. The purpose of this study, then, was to review available universal screening measures used to evaluate SEB constructs of kindergarten readiness that can be used to identify students who could benefit from additional SEB supports.

SEB Constructs of School Readiness

Several components of SEB functioning have been consistently identified in the literature as essential for overall healthy development, especially as they relate to academic learning. These components include the ability to form stable, positive relationships with others, to regulate and express emotions in culturally and socially appropriate ways, and to demonstrate strong self-regulatory behaviors (Alfonso & Flanagan, 2009; Collaborative for Academic, Social, and Emotional Learning [CASEL], 2018; Halle & Darling-Churchill, 2016; Kagan et al., 1995; NEGP, 1997; Yates et al., 2008). For the purposes of this study, SEB functioning at kindergarten entry was

conceptualized as a multi-faceted construct related only to the SEB components of school readiness and excluded the cognitive components of SEB as this study will focus on reviewing indirect assessments of SEB functioning (Alfonso & Flanagan, 2009). SEB functioning, then, was defined to include students' emotional self-regulation, behavioral self-regulation, and social and interpersonal skills. Each component is described below within the context of classroom expectations.

The first component, emotional self-regulation is a sub-component of self-regulation and encompasses a set of skills that students need to control their emotions (Edossa, Schroeders, Weinert, & Artelt, 2018). These set of skills require students to first assess the environment or situation that they encounter and identify and subsequently enact the appropriate emotional response (Raver, Garner, & Smith-Donald, 2007; Ray & Smith, 2010). Children with high sense of emotional self-regulation are better able to inhibit negative emotional responses and initiate more appropriate responses, which allows for better regulation of learning behaviors (Howse, Calkins, Anastopoulos, Keane, & Shelton, 2003; Raver et al., 2007; Trentacosta & Izard, 2007). Conversely, students with poor emotional regulation may demonstrate high levels of externalizing behaviors, such as aggressive behaviors, which are particularly disruptive within the context of a classroom (Barth, Dunlap, Dane, Lochman, & Wells, 2004; Roll, Koglin, & Petermann, 2012). This then highlights the importance of early identification efforts for students with emotional self-regulation difficulties.

Another critical SEB construct of kindergarten readiness and sub-component to self-regulation is behavioral self-regulation (Edossa et al., 2018). Behavioral self-regulatory skills are behaviors related to engagement or learning, and include the ability

to suppress dominant behaviors and enact subdominant less-desirable responses such as following directions and taking turns during group activities (Blair, 2002; Blair & Razza, 2007; Bodrova & Leong, 2008; Linder, Ramey, & Zambak, 2013). Students who display strong behavioral self-regulatory skills are better able to engage with classroom material, leading to greater academic achievement (McClelland & Cameron, 2011). Behavioral self-regulatory skills, along with emotional self-regulatory skills, appear to positively predict future academic achievement (Cooper & Farran, 1988; Howse et al., 2003).

Lastly, social and interpersonal skills essential for kindergarten entry include skills that lead to successful interaction with others in the classroom, such as communication skills, cooperation, and respect for others (CASEL, 2018; DiPerna, Volpe, & Eliott, 2002; Ray & Smith, 2010). These skills are essential for forming strong, healthy, and supportive relationships with both peers and adults in the classroom that lead to positive experiences in the classroom (DiPerna et al., 2002). Further, these skills may also increase students' willingness to engage with the academic material, thereby strengthening their academic learning (DiPerna et al., 2002).

Altogether, these SEB constructs collaboratively work to promote academic achievement in school. Specifically, behavioral self-regulatory skills appear to act as a mediator between emotional regulation and academic success (Edossa et al., 2018; Howse et al., 2003; Raver et al., 2007). Students who are better able to regulate their emotions demonstrate stronger behavioral self-regulatory skills, which may be predictive of prosocial encounters with peers and teachers (Edossa et al., 2018; McClelland & Cameron, 2011). These skills may increase the likelihood of prosocial interactions with peers and teachers, ultimately influencing academic achievement (Howse, Lange, Farran,

& Boyles, 2010; Lopes, Salovey, Cote, & Beers, 2005). Conversely, students who display poor emotional self-regulation and high levels of aggressive behaviors demonstrated poor behavioral self-regulatory skills (Essex et al., 2009; Howse et al., 2003; Raaijmakers et al., 2008). These behaviors may then lead to antisocial behaviors, which reduce the likelihood students will make meaningful connections with others or obtain help and encouragement from others (Coolahan, Fantuzzo, Mendez, & McDermott, 2000). Together, emotional and behavioral self-regulatory skills and social and interpersonal skills are strongly related to future academic achievement (Howse et al., 2003; Li-Grining, Vortruba-Drzal, Maldonado-Carreno, & Haas, 2010; Liew, 2012). Therefore, universal screening for SEB functioning at kindergarten entry is warranted.

Considerations in screening for school readiness

Best practice in universal screening entails consideration of the whole child, which translates to the screening of both academic and SEB difficulties for all students (Glover & Albers, 2007; Ikeda et al., 2008). Thus, the purpose of universal screening is to prevent the onset of later academic problems by identifying students who are experiencing academic and SEB difficulties and in need of additional services, thus increasing the likelihood for future academic success (Essex et al., 2009; Glover & Albers, 2007; Ikeda et al., 2007). Because screening must be completed across multiple timepoints throughout the year, screening procedures need to be relatively low-cost, and require minimal time for administration, scoring, and interpretation (Glover & Albers, 2007; Ikeda et al., 2008). Lastly, and most importantly, universal screening procedures need to accurately and reliably identify students who are at-risk for behavioral concerns and future academic failure (Ikeda et al., 2007).

Unfortunately, less than 15% of schools engage in SEB screening, compared to approximately 80% of schools in the survey who completed yearly academic screening (Bruhn, Woods, Groves, & Huddle, 2014). Barriers to SEB screening included a general lack of awareness on the existence of screeners for such purposes, lack of consensus on which SEB screeners to implement within a school, and a lack of resources (Bruhn et al., 2014, Carter, Briggs-Gowan, & Davis, 2004).

Schools not utilizing systematic screening procedures of SEB functioning are relying on other identification methods, such as the use of existing data (e.g., office discipline referrals, attendance) and teacher and staff reports (Bruhn et al., 2014; Miller et al., 2015; Severson et al., 2007). However, these methods are not as reliable in identifying students with SEB problems as compared to rating scales that have reliability and validity evidence supporting their use (Eklund et al., 2009; Miller et al., 2015). Therefore, it is important that schools have methods to identify the best screening procedures for SEB problems that include the use of rating scales with reliability and validity evidence.

Prior to identifying a specific rating scale for universal screening procedures, a thorough review of psychometric properties is required that entails a review of the reliability, validity, bias identification methods, and norming procedures (Glover & Albers, 2007; Salvia & Ysseldyke, 1998). Strong reliability and validity evidence indicate a rating scale accurately measures the constructs intended to be measured (Salvia & Ysseldyke, 1998). In addition, a well-represented norming sample representative of the general population, as well as thorough bias evaluation procedures, allow for valid interpretation of the results for all students (Glover & Albers, 2007). Therefore, a

thorough review of norming procedures and bias evaluation procedures are warranted, particularly because school readiness characteristics may differ for students from various backgrounds.

Student factors that impact SEB screening accuracy. There are several student factors that may impact ratings of SEB difficulties. Specifically, male students are more likely to display lower school readiness skills and increased behavioral problems (Essex et al., 2009; Montes, Lotcyzewski, Halterman, & Hightower, 2012; Division of Early Childhood Development, 2016) as compared to female students. In addition, students of diverse backgrounds have also demonstrated increased problem behaviors (Montes et al., 2012) and reduced school readiness skills (Division of Early Childhood Development, 2016; Mollborn, 2016).

Parental demographic factors such as income and education-level are also important factors when considering SEB constructs of kindergarten readiness (Mollborn, 2016; Montes et al., 2016). Furthermore, students who reside in the Western region of the United States displayed reduced school readiness skills as compared to students who reside in other regions of the United States (Division of Early Childhood Development, 2016). All of this highlights the importance of reviewing both the norming samples and bias evaluation procedures of rating scales for SEB constructs of school readiness when applicable, particularly to ensure all students are adequately represented in the sample.

Current status of the literature. Given the centrality of these issues to student success, researchers have assembled and reviewed many rating scales that can be used for the purposes of screening for SEB difficulties for students of all ages (Halle & Darling-Churchill, 2016; Jenkins et al., 2014; Severson et al., 2007; Yates et al., 2008). In

addition, several compendia provide lists and reviews of available screeners for use with the kindergarten population (Brown, Scott-Little, Amwake, & Wynn, 2007; Niemeyer & Scott-Little, 2002). However, limited guidance is available for schools as they work to identify most appropriate screeners for use, particularly given the complexity with how the SEB constructs are conceptualized and how they are represented on rating scales (Snow, 2006). In addition, systematic reviews of scales directly related to the SEB components of kindergarten readiness, specifically, are scant, and many do not include an exhaustive review of available SEB rating scales, and instead review a handful of popular rating scales (Carter et al., 2004; Halle & Darling-Churchill, 2016; Jenkins et al., 2014; Oakes, Lane, & Ennis, 2016; Oakes, Lane, Cox, & Messenger, 2014). This is particularly problematic given the importance of these skills to student success. Lastly, researchers may not consistently include cultural bias reviews of scales which demonstrates a scale is appropriate for use with culturally and linguistically diverse students. The current study is a systematic review of universal screeners used to evaluate SEB constructs of kindergarten readiness and includes a thorough review of psychometric properties and bias identification methods. The bias evaluation review is important as it will provide an evaluation of each scale's applicability to diverse student populations, particularly because kindergarten classrooms are becoming increasingly diverse (U.S. Department of Education, 2018). This study can provide guidance to schools as they identify universal screening procedures appropriate to their setting. In addition, the current study will identify areas in which research is lacking and will highlight the need for additional empirical evidence for each scale, and provide direction for future research implications.

Current Study

This systematic literature review aims to identify and review available teacher-reported universal screening measures for the SEB constructs of school readiness in kindergarten. A review of the psychometric properties of each scale will be provided, in addition to a review of the norming procedures used to determine the extent to which each scale is generalizable to students of diverse populations, when appropriate. To this end, this study will answer the following research questions:

1. What are the available universal teacher-report screeners that measure the SEB constructs of school readiness that are currently available for use in a classroom setting for kindergarten students?
2. What are the psychometric characteristics listed for each of the scales?
3. To what extent are the scales appropriate for use with diverse student populations?

Method

This study utilized a systematic review process to identify rating scales that can be used to measure the SEB constructs of kindergarten readiness using the inclusion criteria listed below. To identify relevant rating scales, first, eligible rating scales were located through a review of the Buros Institute of Mental Measurements, starting with the 11th edition published in 1992 to the most recent (20th) edition published in 2017. Next, the most recent School Psychology Best Practices chapters (Albers & Kettler, 2014; Campbell & Hammond, 2014; Dawson, 2014) were reviewed to identify any additional measures not included in the Mental Measurement Yearbook (MMY). Lastly, a review of instruments used in school readiness evaluations published by the Institute of Education Sciences (Brown et al., 2007) and a compendium published by the U.S. Department of

Education on the available assessment instruments for kindergarten children (Niemeyer & Scott-Little, 2002) were reviewed to identify potential measures for inclusion (see Figure 1).

Throughout this process, titles were first screened to exclude rating scales that were clearly measuring constructs other than the SEB functioning of kindergarten students. For example, rating scales that measured student academic achievement, SEB functioning in high school students, or were constructed for the identification of specific disorders or disabilities, such as autism, were excluded from this study. Next, abstracts of rating scales that appeared to measure SEB functioning of kindergarten students were reviewed. Lastly, full review of the rating scale such as those available through the MMY, when available, were reviewed to exclude rating scales that did not meet inclusion criteria for this study.

Inclusion Criteria

Measures were included in this study if they met the following criteria: 1) identified as a rating scale appropriate for universal screening procedures, 2) measured at least one of the SEB constructs of kindergarten readiness as defined by this study, 3) identified a targeted age range that included students of kindergarten age (5 years of age), 4) administered by teachers, 5) was not intended to be used to identify students with specific disabilities or diagnoses (for example, rating scales used to diagnose or identify eligibility for specific disabilities such as autism, ADHD, or students with developmental delay), 6) was not intended for use only with a specific student population (for example, rating scales that measure adaptive skills for students with an already identified disability, such as intellectual disabilities), and 7) developed for use in the United States.

For the purposes of this study, an eligible rating scale was defined as a screening instrument administered in a classroom by a teacher to identify students who may be at risk for or are exhibiting SEB difficulties. Eligible rating scales were required to have the option to be administered universally to all students in a time efficient manner (i.e., universal screening measure). For this reason, rating scales that required more than 10 minutes time to administer per student were removed and deemed ineligible for this study. In addition, rating scales were required to allow for teacher rather than parent administration. Student behavior may differ within a school setting as compared to a home setting; therefore, it is essential teachers complete the rating scale to identify students with SEB difficulties at school (Major & Seabra-Santos, 2015). Examples of eligible scales include the Social Skills Improvement System (SSiS) Performance Screening Guide or the Behavior Assessment System for Children, Third Edition, Behavioral and Emotional Screening System (BASC-3 BESS). Rating scales that were administered for purposes other than screening purposes, such as the BASC-3 Teacher Report Form, were deemed ineligible for this review. Finally, rating scales that were administered for specific diagnostic purposes, such as those administered for the identification of an intellectual disability or attention deficit hyperactivity disorder, were not included in this review.

Additionally, rating scales included in this study were required to provide information on the students' SEB functioning related to kindergarten readiness. For the purpose of this study, the multi-dimensional construct of SEB school readiness was conceptualized as consisting of three key components: behavioral and emotional self-regulation and social and interpersonal skills. However, comprehensive rating scales that

comprehensively measure all constructs of SEB functioning are scarce (Halle & Darling-Churchill, 2016). For this reason, rating scales were included in this review if they measured at least one component of the SEB constructs of kindergarten readiness. Rating scales that meet these criteria provided either a composite score that represents one or more construct of SEB, or provided multiple subscale scores that represent one or more SEB constructs. In addition, rating scales that measured internalizing or externalizing behaviors as a method of screening for problem behaviors were identified as measuring the behavioral and emotional self-regulation constructs.

For scales with multiple editions, only the most recent version of the rating scale was included in this review. In addition, rating scales were deemed eligible if the most current version was published in or after the year 2000 in order to ensure the scales were applicable to current kindergarten students (Floyd et al., 2015). Based on these criteria, a total of 11 rating scales were included for review in this study (Figure 1). Included ratings scales consisted of the following: Behavior Assessment System for Children, Third Edition Behavioral and Emotional Screening System (BACS-3 BESS; Kamphaus & Reynolds, 2007), Behavioral and Emotional Rating Scale, Second Edition (BERS2; Epstein, 1998), Behavior Rating Inventory of Executive Function, Second Edition Screening Teacher Form (BRIEF2 Screening Teacher Form; Gioia, Isquith, Guy, & Kenworthy, 2015), Conners Early Childhood Behavior Short Form (Conners EC(S); Conners, 2009), Children's Aggression Scale (CAS; Halperin & McKay, 2008), Devereux Student Strengths Assessment (DESSA-Mini; Naglieri, LeBuffe, & Shapiro, 2011).), Emotional and Behavioral Screener (EBS; Cullinan & Epstein, 2013), Systematic Screening for Behavior Disorders, Second Edition (SSBD2; Walker,

Severson, & Feil, 2014), School Social Behavior Scales, Second Edition (SSBS-2; Merrell, 2002), Social Skills Improvement System Performance Screening Guide (SSiS Performance Screening Guide; Gresham & Elliott, 2008), and Teacher-Child Rating Scale 2.1 (T-CRS 2.1; Hightower & Perkins, 2010).

Data Extraction

The coding procedures used in this study were adapted based on a previous review of adaptive rating scales conducted by Floyd et al. (2015). Data were extracted through a review of information reported in the assessment manual by the developers. Each scale included in this study was coded to report the following descriptive information, psychometric properties, and bias evaluation.

Descriptive Information. First, general descriptive information for each scale was coded. Specifically, the type of scale (e.g., criterion- vs. norm-referenced) and the age or grade range were coded. In addition, the purpose of the rating scale and the administration time were also noted. Finally, the types of scores used to interpret the results were noted.

Psychometric Properties. Next, psychometric properties were coded for each included rating scale. The psychometric properties coded included the scales' reliability and validity measures, as well as descriptions of the norming samples and reported bias identification measures. Whenever available, the internal consistency, test-retest, and inter-rater reliability results were coded for both entire sample as well as the sub-sample specific kindergarten-aged students.

Internal consistency. Internal correlation or consistency refers to the extent to which multiple similar items on a rating scale are answered consistently by the same

rater. Internal consistency is used to assess the consistency of the results across items within the scale. Internal consistency is typically calculated through formulas or analyses that report correlations such as Cronbach's coefficient alpha or Kuder-Richardson formula 20 (KR_{20}). For the purposes of this study, internal consistency was deemed strong if the reported internal consistency coefficient was measured to be equal to or greater than 0.90, adequate if measured to be between 0.80-0.90, and inadequate if measured to be less than 0.80 for the total composite score across all age or grade range and for specified social-emotional and behavioral domains if applicable (Salvia & Ysseldyke, 1998). The reported internal consistency coefficient for the individual age (includes students 5 years of age) or grade range specific to kindergarten students was deemed strong if the reported internal consistency coefficient was measured to be equal to or greater than 0.90, adequate if it were measured between 0.80-0.90, and inadequate at 0.80 or below (Alfonso & Flanagan, 2009; Bracken, 1987; Floyd et al., 2015).

***Test-retest reliability*.** Test-retest reliability provides a measure of the stability or consistency of the scale's score over a specified period of time. Greater variation between the results of the repeated administrations would indicate greater error in the test scores and lower reliability measures (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). Reports of correlations for test-retest reliability include Pearson product moment correlation coefficients, intraclass correlation coefficients, and Cohen's *d*, for example. Test-retest reliability coefficients were deemed strong if the coefficients were greater than 0.90 and adequate if the coefficients were between 0.80 and 0.90 for the total composite scores. The test-retest reliability coefficients were deemed adequate if

measured to be 0.80 or greater for specific domains or skills areas (Alfonso & Flanagan, 2009; Bracken, 1987; Floyd et al., 2015). The test-retest reliability coefficients for the specified age or grade range specific to kindergarten students were also reported (Floyd et al., 2015). Test-retest reliability coefficients were deemed inadequate for any section listed above if measured to be less than 0.80. The time period between the first and second administration was also coded.

Interrater reliability. Interrater reliability assesses the consistency of scale scores as completed by different raters to ensure consistency in the administration process (Salvia & Ysseldyke, 1998). Interrater reliability coefficients can be reported as correlations such as Pearson product moment correlation coefficients, intraclass correlation coefficients, or Cohen's *d*. Interrater reliability coefficients were deemed adequate if the reported interrater reliability coefficient was equal to or greater than 0.60 for the total composite and SEB domain, if applicable (Floyd et al., 2015). Additionally, reliability coefficients broken down by age or grade ranges were reported specifically for the age or grade group that included kindergarten students (Floyd et al., 2015) and were rated in a similar manner to the composite and SEB domain listed above.

Validity. Validity of the scales was evaluated as evidence based on the following five categories: 1) the scale content, 2) the internal structure, 3) the external relations with other variables, 4) the response process, and 5) the consequences of testing (AERA et al., 2014; Floyd et al., 2015).

Evidence based on content demonstrates the extent to which the scale's content represent the constructs in an accurate, complete, and unbiased manner. Scale content may include the characteristics of the items, such as the wording and format, as well as

the administration and scoring procedures of the scale. This type of validity evidence provides assurance the constructs within the scales are the ones actually being measured. Examples of adequate evidence based on content include citations of relevant work or evidence from subject matter experts indicating the items used in the scale fit onto the proposed interpretation of the scale.

Next, evidence based on internal structure of the scale provides information on the extent to which the test items correlate with the constructs of the proposed interpretation of the scale. Measurement of the internal structure of a scale is dependent on the proposed use of the scale. Evidence based on internal structure can be demonstrated through correlational evidence used to examine the extent to which different items represent a similar construct through exploratory factor analysis or confirmatory factor analyses (Rios & Wells, 2014).

The third type of evidence, evidence related to external variables, was included to identify systemic relationships between the scale's score and external characteristics. External characteristics can include other scales intended to measure different or similar constructs, or categorical variables such as group memberships (e.g., the identification of group differences in the sample based on diagnoses or special education identification). This type of evidence can be used to determine criterion-related, concurrent, and divergent validity. Evidence of this type of validity can be obtained through experimental or correlational evidence.

The fourth type of evidence, evidence based on the response process provides evidence based on the assumptions made by the rater as they complete the rating scale. This type of evidence provides assurance that the rater is indeed responding to the items

as expected, and is usually presented as interviews with raters to track their thoughts or response processes. Lastly, evidence based on the consequences of testing provided evidence the scores and the decisions made based on the scores were accurate and produced intended consequences for the students being assessed. This type of evidence can be measured through an evaluation of treatment validity, which identifies any positive changes following intervention implementation based on results of the assessment (AERA et al., 2014).

Overall, for the purposes of this study, adequate validity evidence was defined as the identification of at least five sources of evidence within each of the five identified validity evidence types. Good validity evidence was defined as having three to four sources of evidence. Lastly, an inadequate evidence rating was assigned if the scale presented less than three sources of evidence (Alfonso & Flanagan, 2009; Bracken, 1987; Floyd et al., 2015, Halle & Darling-Churchill, 2016).

Norming sample. Data collected on the norming sample pertains solely to norm-referenced rating scales. A total of 10 out of the total 11 rating scales were identified as norm-referenced scales. Data collected included information on the norming sample size, date of data collection, geographic locations of data collection, subsample characteristics if applicable, and information on U.S. census data. The quality of the norming was evaluated based on the date of collection (strong if collected within the past 10 years, adequate if collected within the past 15 years, and inadequate if collected more than 15 years ago; Alfonso & Flanagan, 2009; Floyd et al., 2015), geographic locations (strong if data were collected from more than 35 states, adequate if collected from 25 to 35 states, and inadequate if collected from less than 25 states; Floyd et al., 2015), and total sample

size (strong if the sample size is greater than 100, adequate if between 30-99, and inadequate if less than 25; Floyd et al., 2015). The sample size of the group that included kindergarten age was also evaluated. In addition, any great differences between the characteristics of the norming sample, such as community size or socio-economic status, and the U.S. Census Data were noted.

***Bias identification*.** The purpose of bias identification was to determine if the scale is biased in its administration or scoring procedures towards any specific group. These errors are systematic in nature rather than random error that may occur through the administration and scoring procedures (Reynolds & Carson, 2005). Differential item functioning analyses were reported to indicate the extent to which raters of similar groups rated similar items differently. Evaluation of bias identification was identified as good if five or more evaluation methods were reported, adequate if three or four methods were reported, and inadequate if less than three methods were reported (Floyd et al., 2015).

Results

Scale Description

A total of 11 rating scales met this study's eligibility criteria for review. Table 1 provides a description of each scales' basic characteristics. Most scales provided a description of the scale's purpose that allowed for immediate implementation of intervention services as needed, without the need for additional assessment (e.g., BASC-3 BESS, SSBD2). Other scales' descriptions specifically stated a need for further assessment prior to intervention implementation (e.g., BRIEF2 Screening Teacher Form). Two of the included rating scales' manuals indicated a purpose related to identifying students at-risk for emotional disturbance specifically (BERS2, EBS).

Six of the eleven rating scales yielded single composite scores that summarize the students' overall SEB skill level, such as the BASC-3 BESS and the BRIEF2 Screening Teacher Form. The remaining scales provided scores for several primary domains, such as the SSiS Performance Screening Guide's scores for Reading Skills, Math Skills, Prosocial Behavior, and Motivation to Learn. Most rating scales provided scores for more specific secondary domains. For example, the T-CRS 2.1 provided eight additional secondary subscale scores in addition to the composite scores, such as students' Learning Skills and Frustration Tolerance. These secondary scales provide more specific information regarding the students' functioning. Given that psychometric properties are strongest at the broadest level (e.g., total reported score or primary domain scores) as compared to subscale measures (Monaghan, 2006), psychometric properties were coded only at the broadest level available.

All identified rating scale manuals require teachers to complete the scales for all students, with the exception of the SSBD2. The procedures outlined in the SSBD2 manual requires teachers to utilize a multi-gated process for identifying students in need of SEB interventions. First, teachers must list and rank students with highest behavioral needs for internalizing and externalizing behaviors separately (Stage 1). Next, teachers complete several rating scales for the three students with highest internalizing and externalizing needs separately (Stage 2). Because the SSBD2 utilizes a multi-gated process, the psychometric properties of this scale were coded for each stage separately as appropriate.

Psychometric Properties of Scales

Overall, reliability and validity data appeared lacking for several rating scales. Several rating scale manuals failed to report reliability or validity evidence for the entire sample or for kindergarten-aged students. Furthermore, several manuals reported reliability or validity evidence only for students in specific grades or ages. Table 2 provides a summary of the psychometric properties of each rating scale, including a summary of reliability, validity, bias evaluation, and norming sample characteristics.

Reliability. In order to determine the screeners' reliability, each rating scale manual was reviewed to identify information on internal consistency, test-retest reliability, and interrater reliability. Various rating scale manuals reported reliability evidence as measured for the entire study sample, as separated by age or grade groups, or both. Results provided here will include a review of reliability evidence for the entire sample as well as separated by age or grade groups, as applicable. In addition, several rating scale manuals reported reliability evidence for samples representing the general population and clinical populations. In such cases, reliability evidence was reported for all reported samples.

Internal consistency. Internal consistency was identified as strong if the correlation coefficient was greater than 0.90, adequate if the correlation coefficient was between 0.80 and 0.90, and inadequate if the correlation coefficient was less than 0.80. Two rating scale manuals did not report internal consistency (SSBD2 and SSiS Performance Screening Guide). The SSBD2 manual provided internal consistency for students in grades other than kindergarten. The SSiS Performance Screening Guide did not provide internal consistency evidence specifically relating to the screening form.

The remaining rating scale manuals reported internal consistency evidence for the entire sample, with the exception of the BASC-3 BESS manual, which reported internal consistency by age groups only. Most manuals reported either strong or adequate internal consistency. Inadequate internal consistency for the entire sample was identified for the Anxiety and Mood and Affect domains of the Conners EC(S). When examined by age groups, internal consistency was reported for five of the eleven manuals reviewed (BASC-3 BESS, BERS2, BRIEF2 Screening Teacher Form, EBS, and SSBS-2). Internal consistency reported by age group were strong or adequate, with the exception of the EBS, and only relating to the sample of students without emotional disturbance.

Test-retest reliability. Test-retest reliability was deemed to be strong if the correlation coefficient was measured to be equal or greater than 0.90, adequate if measured to be between 0.80 and 0.90, and inadequate if measured to be less than 0.80. The test-retest reliability evidence reported in the SSBS-2 manual pertained to students in grades other than kindergarten and not included in this review. Of the remaining manuals, all provided test-retest evidence for the entire sample. None of the ratings scales reported test-retest reliability for age-specific groups.

The test-retest reliability evidence for the SSBD2 was reported for each stage separately. In addition, the BERS-2 manual documented two separate studies that were conducted to measure test-retest reliability, one with students without emotional and behavioral disabilities (EBD), and one with students at-risk for EBD. Test-retest reliabilities were measured using subsamples that ranged from 28 to 302 participants. Time intervals between the first and second administration of each rating scale ranged from 4 days to 7 months. Test-retest reliability evidence was coded as strong or adequate

for all scales, with the exception of three: Stage 1 and 2 of the SSBD2, the SSiS Performance Screening Guide, and the T-CRS 2.1.

Interrater reliability. Interrater reliability of scales was rated as adequate if the correlation coefficient was measured to be greater than 0.60. Interrater reliability was assessed using Pearson correlation coefficients on all scales with the exception of the SSiS Performance Screening Guide and the SSBD2 manuals which yielded intraclass correlation coefficients and Kappa coefficients, respectively. Five ratings scale manuals did not provide applicable interrater reliability evidence. Specifically, the Conners EC(S) and the T-CRS 2.1 manuals did not report any results pertaining to interrater reliability, and the BERS2, EBS, and SSBS-2 manuals summarized interrater reliability studies with students in grades other than kindergarten. In addition, the BRIEF2 Screening Teacher Form manual provided interrater reliability evidence for a subsample of the standardization sample and the clinical sample separately. Lastly, none of the ratings scales reported interrater reliability for age-specific groups.

Sample sizes of the remaining six rating scale manuals ranged from 51 to 583. Results of the interrater reliability studies indicated inadequate ratings for both student samples of the BRIEF2 Screening Teacher Form, Stage 1 of the SSBD2, and the Prosocial Behavior domain of the SSiS Performance Screening Guide. Interrater reliability was deemed adequate for the BASC-3 BESS, CAS, DESSA-Mini, Stage 2 of the SSBD2, and the Motivation to Learn domain of the SSiS Performance Screening Guide rating scales.

Validity. Validity evidence was reviewed based on the following evidence types: 1) content, 2) internal structure, 3) external relations with other variables, 4) response

process, and 5) consequences of testing (AERA et al., 2014; Floyd et al., 2015). Validity of rating scales were coded as strong if manuals identified at least five sources of evidence for each of the evidence type mentioned here, adequate if manuals identified three to four sources of evidence, and inadequate if manuals identified less than three sources of evidence (Floyd et al., 2015). Overall, validity evidence appeared lacking. None of the rating scales demonstrated evidence based on all five types. Of the validity evidence provided, most were rated inadequate as they demonstrated one or two sources within each type.

All rating scale manuals provided validity evidence based on content. Specifically, the BERS-2, Conners EC(S), SSB2D, SSBS-2, and T-CRS 2.1 demonstrated strong evidence based on content, and the EBS manual presented adequate evidence. The remaining scale presented less than two sources of evidence based on content. Three of the eleven scale manuals did not report evidence based on internal structure (Conners EC(S), SSBD2, and SSiS Performance Screening Guide). Of the remaining scales, all manuals reported less than three sources of evidence based on internal structure with the exception of the BERS2 and the SSBS-2 manuals, which provided adequate sources of evidence. With regard to evidence based on external relations, all scale manuals provided criterion-related evidence, and nine manuals provided external relations evidence based on group-differences. Four manuals reported strong criterion-related evidence (BASC-3 BESS, BRIEF2 Screening Teacher Form, CAS and Conners EC(S)), and two reported adequate evidence (BERS2 and SSBS-2). All rating scale manuals provided inadequate evidence related to group differences with the exception of the SSBD2 and the T-CRS 2.1, which provided adequate sources of evidence. Next, only two scale manuals

(DESSA-Mini and SSBD) reported any evidence based on the consequences of testing, with an inadequate rating. Lastly, none of the scales demonstrated evidence based on the response process.

Norming Samples. The quality of the norming was evaluated based on the sample size specific to kindergarten students (strong if the sample size is greater than 100, adequate if between 30-99, and inadequate if less than 25; Alfonso & Flanagan, 2009; Floyd et al., 2015), the date of collection (strong if collected within the past 10 years, adequate if collected within the past 15 years, and inadequate if collected more than 15 years ago; Floyd et al., 2015), and geographic locations (strong if data were collected from more than 35 states, adequate if collected from 25 to 35 states, and inadequate if collected from less than 25 states; Floyd et al., 2015). All rating scales included were norm-referenced with the exception of the SSiS Performance Screening Guide; therefore, the SSiS Performance Screening Guide was not included in this part of the review.

The rating scales' total sample size ranged from 700 (T-CRS 2.1) to 3037 (BERS-2). The sample size specific to the age interval which included kindergarten age ranged from 71 (EBS) to 915 (Stage 1 of SSBD2). Sample sizes of the EBS and the SSBS-2 were rated as adequate. All other rating scale manuals reported strong sample sizes. The BERS-2, CAS, Conners EC(S), and EBS scales intentionally included students from clinical populations in their norming samples (e.g., with identified disabilities or clinical diagnoses).

With regards to recruitment dates, two rating scale manuals did not provide dates of data collection (BRIEF2 Screening Teacher Form, EBS). Of the remaining scales,

dates of data collection ranged from years 1990 (SSBS-2) to 2013 (BASC-3 BESS). The BASC-3 BESS, CAS, DESSA-mini, and EBS demonstrated strong or adequate ratings of time of data collections. The remaining scales were deemed inadequate as their norming data was collected more than 15 years ago. Surprisingly, the norming sample of the SSBS-2 and the SSBD2 included norming data initially used in the first editions of the scales, however, the SSBS-2 manual also indicated data collection for additional norming data in 1998.

Furthermore, the geographic locations of data collection for most rating scale manuals were rated to be either adequate or strong. Only three rating scales reported the occurrence of data collection in less than 25 states (SSBD2, SSBS-2, and T-CRS 2.1). In addition, no demographic data were documented for the CAS, Conners EC B(S), and the DESSA-Mini.

Finally, all rating scales manuals provided information on the comparisons of the norming sample descriptive information to Census data, with the exception of the SSBD2 and T-CRS 2.1. Each rating scale reported descriptive data in different manners. For example, the BERS-2 manual provided descriptive information for the students with and without an ED separately, and the BRIEF2 Screening Teacher Form manual separated all descriptive data by gender. In addition, most manuals reported parental income or free and reduced lunch (FRL) status as descriptors of parental socioeconomic status (SES).

Bias evaluation. Ratings of bias evaluation was determined by the total number of evaluation methods reported within each scale manual. Evaluation of bias identification was identified as good if five or more evaluation methods were reported, adequate if three or four methods were reported, and inadequate if less than three

methods were reported (Floyd et al., 2015). Four of the eleven rating scales manuals did not evaluate any potential bias in administration or scoring procedures (BASC-3 BESS, BRIEF2 Screening Teacher Form, SSBD2, and SSiS Performance Screening Guide). Of the remaining rating scales, only one manual demonstrated strong evidence for bias evaluation (T-CRS 2.1) and three demonstrated adequate evidence (Conners EC(S), EBS, and SSBS-2). All other ratings scale manuals were identified as having inadequate bias evaluation.

Finally, the demographic characteristics that were evaluated as part of the bias evaluation were noted. Seven total evaluations of gender differences were conducted (CAS, Conners EC(S), EBS, SSBS-2, and TCRS-2.1), and five evaluations of differences based on student race were reported (BERS2, Conners EC(S), DESSA-Mini, EBS, and TCRS-2.1). Only the T-CRS-2.1 also included bias evaluation results based on student SES, region of residence, and grade level.

Discussion

This systematic review aimed to identify and review available universal rating scales that measure the SEB constructs of school readiness that are available for use in a kindergarten classroom setting. This review also aimed to evaluate the extent to which each scale is appropriate for use with diverse student populations. To this end, this study identified and reviewed eleven rating scales that can act as universal screeners for the SEB constructs of school readiness in a kindergarten classroom. Each identified rating scale manual identified screening as a purpose for administration, albeit for different outcomes. Some rating scale manuals identified intervention planning as a final outcome of scale completion, whereas others sought to identify students who require additional,

more detailed evaluation to identify SEB deficits, or to identify students who were at risk for specific disabilities, such as emotional disturbance. The first type of rating scales appear to be most useful within a multi-tiered systems of support (MTSS) framework as they work to identify students in need of additional intervention services in order to strengthen their SEB functioning in the classroom (Ikeda et al., 2008). The identification of an appropriate rating scale that can be used within the framework of MTSS is particularly important for the success of intervention implementation (Parisi, Ihlo, & Glover, 2014). School personnel must ensure that rating scales' purpose for administration matches the ways in which the rating scales will be used within their school setting (Parisi et al., 2014). In addition, it is important to note that the identified rating scale will guide future decisions, highlighting the importance of identifying the most appropriate rating scale for each setting (Kamphaus, Reynolds, & Dever, 2014). By appropriately using rating scales within the context of universal screening, students at-risk for future behavioral and academic difficulties due to current SEB concerns can be identified for early intervention services that can reduce or deter future difficulties (DiPerna, Bailey, & Anthony, 2014).

In addition, all included rating scales were required to measure core constructs of SEB readiness, defined here as students' ability to regulate their emotions, demonstrate strong behavioral self-regulatory skills, and interact with others in positive and culturally appropriate ways (CASEL, 2018; Halle & Darling-Churchill, 2016; NEGP, 1997; Yates et al., 2008). Slight variations in the measurement of SEB constructs, however, were noted. The outcome measure of most rating scale was one composite score that summarized the students' overall SEB skills, whereas others also provided more detailed

subscale scores that measured the SEB constructs separately. Alternatively, several rating scales focused on the measurement of just one or two constructs of SEB functioning rather than overall SEB functioning in the classroom. This inconsistency in the measurement of SEB functioning may make it difficult to identify appropriate rating scales for use within a universal screening framework. In addition, past reviews of SEB constructs have noted that operationalized definitions of SEB constructs within rating scales are lacking, and becomes particularly more difficult when considering specific subdomains of SEB functioning (Halle & Darling-Churchill, 2016).

Lastly, it should be noted that some rating scale scores represented students' SEB strengths rather than deficits. Strength-based assessments allow teachers and students to focus on and celebrate the students' strengths, and replace students' deficits with opportunities for learning and development (Epstein, 1998; Epstein et al., 2003; Merrell, 2010). The strength-based rather than deficit-based approach can be used to develop intervention supports that capitalize on student strength (Carter et al., 2004). It may also lead to greater parental acceptance and involvement with intervention implementation (Epstein et al., 2003), which may then translate to greater academic success (Fan & Chen, 2001). Ultimately, however, these drastic differences in the methods available for measuring SEB constructs makes it difficult for school personnel to identify the appropriate screeners for use (Carter et al., 2004).

The second aim of this study sought to review the identified rating scales' psychometric properties. Overall, this study determined that a large majority of the rating scale manuals provided strong or adequate reliability evidence for all reliability types as measured by the overall student sample, and for internal consistency measure for

kindergarten students specifically. Several factors may influence reliability estimates, including the length of the rating scale and the test-retest interval (Salvia & Ysseldyke, 1998). Longer rating scales will often result in stronger reliability estimates. However, a key feature of universal screening measures involves the use of short measures that are time- and resource-efficient (Glover & Albers, 2007; Ikeda et al., 2008). It is essential, then, that rating scale manuals identify the best combination between scale items and reliability estimates. In addition, shorter test-retest intervals represent stable measures and are associated with stronger reliability estimates, highlighting the importance of consistency in the measurement of test-retest reliability across rating scales.

It is also important to note that test-retest and interrater reliabilities were not measured specifically for kindergarten students for all rating scales, a measurement of reliability recommended by Anastasi and Urbina (1997). Ensuring reliability in rating scale administration across raters and over-time may also lead to reduced reliability estimates, particularly when considering the fast-paced changes in students' behaviors that can occur during early childhood (Nagle, 2000). Overall, omitting reliability estimates specific to this age group makes it difficult to determine if such rating scales are appropriate for use within the kindergarten population (Alfonso & Flanagan, 2009).

Overall, reporting of validity evidence appeared to be inconsistent across manuals. In this study, validity evidence was rated based on the number of sources reported to assess each of the five types of validity put forth by AERA et al. (2014; Alfonso & Flanagan, 2009; Bracken, 1987; Floyd et al., 2015; Halle & Darling-Churchill, 2016). This type of evaluation was conducted as validity evidence tends to vary greatly across rating scale manuals, making it difficult to provide criteria by which to evaluate

validity evidence (Alfonso & Flanagan, 2009). For this reason, many reviews of rating scales, including this review, have adopted this procedure for evaluating validity evidence (e.g., Bracken, 1987; Floyd et al., 2015). However, this inconsistency in reporting of validity evidence also makes meaningful evaluation and comparisons of the quality of the evidence across rating scale manuals difficult (Alfonso & Flanagan, 2009).

This study determined that most rating scale manuals demonstrated inadequate validity evidence overall as evidenced by the limited number of sources reported for each type of validity. Most rating scale manuals identified one or two sources of evidence for validity evidence based on content, internal structure, and external relations. Only two rating scale manuals presented any evidence based on consequences of testing, and none of the manuals reported any evidence based on the response process. Validity based on the response process provides support for the interpretation of the administration procedures and completion of the items on the scale (AERA et al., 2014). Without it, one cannot be confident that the rating scale items are written in such a manner that all administrators are interpreting them in a similar fashion (AREA et al., 2014). In addition, validity based on consequences of testing is particularly important in the context of universal screening. One important purpose of conducting universal screening allows for informing the intervention planning process, meaning that the rating scale manual specifically identifies intervention planning as a valid, possible outcome following the scale administration (Cook, Volpe, & Livanis, 2010; Glover & Albers, 2007). Strong validity evidence based on consequences of testing would provide evidence that support the intervention decisions made following the completion of the rating scales (AREA et al., 2014). Although all rating scales identified screening procedures as a primary purpose

for administration, they lacked evidence to support its use as part of the intervention process. This lack of evidence based on consequences of testing is unsurprising, and has been noted as an area in need of great improvement (Cook et al., 2010; Volpe, Briesch, & Chafouleas, 2010). Nevertheless, when used correctly, SEB rating scales can provide valuable information regarding the students' current and future behavioral and academic functioning (Snow, 2006).

Lastly, review of norming data resulted in good ratings solely for sample size. The recency of data and national sampling criteria were rated poorly, suggesting that the norming data provided may not be representative of the current population, and may benefit from newer, more diverse samples. This is particularly important as it may then render the results of the scales invalid. For this reason, best practice guidelines recommend that norming samples be updated every 10 years or less (Alfonso & Flanagan, 2009).

Overall, the inconsistency in evaluating and reporting psychometric properties identified by this study may make it difficult for inexperienced consumers, such as school personnel, to correctly interpret reported psychometric properties and identify reliable and valid rating scales for use. Although teachers commonly state that SEB functioning is an important marker for kindergarten readiness (Lin et al., 2003), many schools identified significant barriers for the implementation of screening procedures, including a lack of knowledge in identifying and appropriately using rating scales (Bruhn et al., 2014). For this reason, the use of best practices relating to assessment would result in consistent reports of psychometric properties, such as those outlined by AERA et al. (2014) and Salvia & Ysseldyke (1998).

The third and final aim of this study involved a review of the scales' bias evaluation methods, which was determined to be severely lacking overall by this study. Bias evaluation guidelines typically include evaluations of potential bias for each reliability or validity type described by AERA et al. (2014; Reynolds & Carson, 2005). It was surprising, then, that most ratings scales identified in this study did not complete a thorough evaluation of potential biases in administration or scoring procedures. In fact, none of the rating scales included in this study completed a bias evaluation for each type of reliability and validity evidence. Most bias evaluations conducted in the identified rating scale manuals pertained to validity based on relations to other variables. Specifically, ratings scale manuals reported potential differences in scale scores based on group memberships such as race, socioeconomic status, or grade.

Bias evaluation procedures are essential as they provide a measurement of any biases in the scales' administration or scoring procedures related to the students' demographic characteristics (Reynolds & Carson, 2005). This type of evaluation is a "fundamental validity issue and requires attention throughout all stages of test development and use" (p. 49, AERA et al., 2014). They are particularly important within the context of SEB screening because drastic differences in the rate of problem behavior identification has been documented for students of varying demographic variables, particularly for students of differing race or ethnic backgrounds (Bates & Glick, 2013; Downey & Pribesh, 2004; Pigott & Cowen, 2000). Overall, the reviewed rating scale manuals included in this study did not provide sufficient sources of bias evaluation, indicating that bias evaluation evidence is severely lacking and requires further assessment. Biases inherent to a scale can be harmful if it over- or under-identifies

students of specific demographic groups such as cultural or linguistic background as at-risk for SEB difficulties in kindergarten (AERA et al., 2014). It is important that teachers and school personnel consider the students' diverse demographic characteristics when interpreting the scale results in order to attempt reduction in construct-irrelevant variance (AERA et al., 2014). One way to control for the over- or under-identification of students from varying demographic variables can include the assignment of different cutoff scores used to detect risk based on student- and school-level factors (Cook et al., 2010). Nonetheless, teachers can do this by first acquiring a strong understanding of the scales' items and systematically completing the scales for all students in order to reduce bias throughout the administration and scoring procedures (AERA et al., 2014). This can be difficult as recent experiences, such as a behavioral outburst in the classroom, may influence the teachers' item ratings (Abikoff, Courtney, Pelham, & Koplewicz, 1993; Smith-Millman et al., 2017). Therefore, teachers must take the time to think through and complete each item fairly for all students. Only then can interpretations of students' scores remain valid and result in proper screening procedures.

Limitations and Directions for Future Research

This study has several limitations that must be noted. First, ratings scales identified for this study were identified through a review of available scales listed in the Mental Measurement Yearbooks, applicable Best Practices in School Psychology chapters, and applicable compendiums. Every effort was made to ensure that all appropriate rating scales were included in the study, however, it is possible that rating scales were missed and excluded from this review. Second, rating scales that did not have a readily-available technical manual were not included in this study. It is possible, then,

that applicable rating scales without compiled technical manuals were not identified for review. In addition, the author of this study focused the psychometric review of rating scales based only on evidence provided by the authors rather than other researchers in order to document psychometric properties obtained solely prior to publication and dissemination of the scales. School personnel may not have access to updated psychometric evidence provided through extant literature, nor do they have the time and resources to thoroughly review updated psychometric properties of each rating scale. Therefore, the reliability or validity domains of the reviewed rating scale manuals that lacked psychometric evidence were deemed missing or inadequate. This study represents an important first step by providing a snapshot of the limited evidence that is provided by rating scale manuals, and may highlight the need for thorough evaluation of psychometric evidence prior to the publication and distribution of the rating scale. However, it is possible that more updated psychometric evidence published in peer-reviewed journals was missed. Future research studies should expand the search to include both applicable rating scales without published manuals as well as updated results of validation studies. Furthermore, this study evaluated the number of ways in which validity and bias evaluation evidence was measured without evaluating the quality of the validity evidence reviewed. It is essential that the quality of the validity evidence as well as the quantity of evidence be reviewed by future studies in order to complete a comprehensive assessment of the scales' psychometric properties. Lastly, future reviews of rating scales should include a cost-benefit analysis which can be useful in the context of universal screening in a school setting as resources are limited (Cook et al., 2010).

In addition, this study has identified significant gaps in the current landscape of assessments for universal screening measures. Specifically, this study demonstrated a need for further examination of the psychometric properties of most rating scales reviewed in this study. First, researchers can expand on the psychometric properties by exploring reliability measures specific to kindergarten-aged students. As noted in this study, none of the included rating scales demonstrated interrater or test-retest reliability specifically for students in kindergarten, which brings into question the consistency of the scales' results across raters of kindergarten students and over time. This can prove problematic when universal screening efforts are implemented multiple times per year.

Other research avenues include the examination of the validity and bias evaluation of rating scales included in this review. This study demonstrated a clear lack of evidence for each validity type for most included rating scales. By expanding on current findings, future research studies can further demonstrate the validity of rating scales, ensuring that these scales will accurately identify students in need of additional supports and services. Furthermore, additional bias evaluation studies can provide support for the use of these rating scale with diverse populations. This is particularly important as kindergarten classrooms become increasingly more diverse (U.S. Department of Education, 2018). Overall, these future research efforts can provide additional support for the use of these rating scales as universal screening measures within the kindergarten classroom setting.

Implications for Practice

This study provides support for professionals in school settings in several ways. First, this review can serve as a guide for identifying SEB rating scales for use in

kindergarten classrooms. Universal screening procedures of SEB functioning are seldom completed in schools across the nation, with the most frequently cited reason being that individuals were unaware that such screening tools existed (Bruhn et al., 2014). This review systematically identified available ratings scales that identify students with SEB difficulties in kindergarten through a review of scales included in the MMY, School Psychology Best Practices chapters, and several compendia related to behavioral screening and kindergarten readiness. Rating scales included in this study were purposefully identified with the guidance of best practices procedures related to universal screening, including specific purpose for scale administration, and minimal time for administration, scoring, and interpretation. These criteria lead to the identification of low-resource, time-efficient rating scales that can be completed for all students in a given classroom. School personnel can use this information when identifying rating scales to determine which scales are most appropriate for their setting depending on the school's goal for implementing screening procedures. For example, school personnel in need of screening procedures for identifying students at-risk for emotional disturbance may find benefit in using the EBS as compared to other rating scales.

In addition, this study also provided a review psychometric properties and bias evaluation procedures presented by each identified rating scale manual. Reviewing the psychometric properties of rating scale manual can be a time-intensive and daunting task. School personnel can use this study's findings to review the psychometric properties of any of the included rating scales, alleviating them of the task of reviewing the entire rating scale manual.

Lastly, school personnel may also choose to review the characteristics of the norming samples documented in the rating scale manuals to determine the extent to which the rating scale of choice is appropriate for use in their specific setting. Schools with high proportions of students from diverse racial and ethnic background may consider identifying rating scales with well-represented norming samples and thorough bias evaluation evidence, and may use the results of this study as a guide. In any case, this review can alleviate the time and effort required for identifying SEB screening measures for kindergarten students.

Conclusion

Results of this systematic review highlight the importance for school personnel to fully review scales' technical manual prior to their administration. Most rating scales reviewed in this study appeared to report strong reliability evidence within the manuals overall. However, validity and bias evaluation evidence were severely lacking. This review identified several aspects of each rating scale with inadequate ratings. For this reason, school personnel should take the time to critically review scales and use this information when identifying appropriate rating scales for use in their settings. In addition, future research efforts should be directed towards establishing stronger psychometric evidence for these scales.

Table 1
Description of Scale Characteristics

Scale	Type of scale	Age or grade range	Administration time	Purpose	Primary measures	Scores
Behavior Assessment System for Children-3 Behavioral and Emotional Screening System	Norm-referenced	Preschool to high school	5 minutes per student 15 minutes per group	"[BASC-3 BESS is] used to determine a child's risk level for developing emotional and/or behavioral problems that require intervention" (p.1)	Behavioral and Emotional Risk Index	T-scores; PR; Risk level
Behavioral and Emotional Rating Scale 2	Norm-referenced	5-18 years old	10 minutes per student	"It [BERS-2] has five principal uses: (a) to identify children with limited strengths, (b) to target goals for an IEP or individual treatment plan, (c) to identify strengths and weaknesses for intervention, (d) to document progress in a strength area as a consequence of specialized services, and (e) to measure strengths in research and evaluation projects." (p.8)	Strength Index composed of sum of subscales: Interpersonal strength; Family involvement; Intrapersonal strength; School functioning; Affective strength subscale	ScS; StS; PR
Behavior Rating Inventory of Executive Function, Second Edition Screening Teacher Form	Norm-referenced	5-18 years old	10 minutes per student	"[The BRIEF2 Screening Teacher Form is] designed for use in medical and educational settings to help determine whether more comprehensive assessment is appropriate" (p.81) "[The BRIEF2 Screening Teacher Form is] designed to be sensitive to the presence of executive function difficulties to determine whether there is a need for more in-depth assessment" (p.81)	Executive function screening raw score	PR

Scale	Type of scale	Age or grade range	Administration time	Purpose	Primary measures	Scores
Children's Aggression Scale	Norm-referenced	5 to 18 years old	10 minutes per student	"The CAS can enable more systematic evaluation of the nature, prevalence, and severity of aggressive behaviors within the classroom or on a broader level (e.g., in a school, in a school district) so that relevant interventions or strategies can be implemented to reduce aggressive acts" (p. 8)	Total aggression index	T-scores; PR
Conners Early Childhood Behavior Short Form	Norm-referenced	2 to 6 years old	10 minutes per student	"The Conners EC can be administered in group settings to help identify children who may require further evaluation for unknown or suspected areas of concern" (p. 7)	Inattention/hyperactivity; Defiant/aggressive behaviors; Social functioning/atypical behavior; Anxiety; Mood and affect; Physical symptoms	T-scores; T-score graph; PR
Devereux Student Strengths Assessment – Mini	Norm-referenced	K to 8th grade	1 minute per student	"The DESSA-mini is a technically sound, user-friendly screening and progress-monitoring tool that has been developed to efficiently measure and track a subset of predictors of future mental, emotional, and behavioral disorders in order to make early intervention more possible." (p.6)	Social-emotional total (SET) score	T-scores; PR
Emotional and Behavioral Screener	Norm-referenced	5-0 to 17-11	2 minutes per student	"[The EBS is used] to identify students at risk for emotional and behavioral problems" (p. 1) "[The EBS is used] to identify students between the ages of 5-0 and 17-11 who are at risk of being identified as having emotional disturbance (ED)" (p.1)	Total EBS score	Other (raw score)

Scale	Type of scale	Age or grade range	Administration time	Purpose	Primary measures	Scores
Systematic Screening for Behavior Disorders	Norm-referenced	Pre-K to 9th grade	<1 hour per classroom	"SSBD is an evidence-based screening system for identifying students who are at risk for internalizing and externalizing problems."	Aggression behavior scale; Social interaction scale	StS; PR
School Social Behavior Scales-2	Norm-referenced	5 to 18 years old	8-10 minutes per student	"Screening tools for identifying children and youth who are behaviorally at risk and who may benefit from prevention or intervention efforts" (p.3)	Social competence scale; Antisocial behavior scale	T-scores; PR
Social Skills Improvement System Performance Screening Guide	Criterion-referenced	3 to 18 years old	30 minutes per classroom	"[SSiS Performance Screening Guide is] designed to enable educators to efficiently collect and organize important information about students' academic performance and learning behaviors that influence performance." (p.155)	Prosocial behavior; motivation to learn; reading skills; math skills[a]	Other (performance level)
Teacher-Child Rating Scale 2.1	Norm-referenced	Pre-K to 8th grade	3-5 minutes per student	"As a screening instrument, it can be used to identify children at risk of performing poorly in school due to socio-emotional problems, as well as children who have strong socio-emotional competencies." (p.4)	Task orientation; Behavioral control; Assertiveness; Peer social skills	PR

Note. BRIEF2 = Behavior Rating Inventory of Executive Function, Second Edition; PR = percentile rank; ScS = scaled score; SEB = social-emotional and behavioral; StS = standard score

[a] Psychometric properties reported for prosocial behavior and motivation to learn domains only.

Table 2
Summary of Psychometric Properties

Measurement Property	Sample	BASC-3 BESS	BERS2	BRIEF2 Screening Teacher Form	CAS	Conners EC Behavior Short Form	DESSA-Mini	EBS	SSBD	SSBS-2	SSiS Performance Screening Guide	T-CRS 2.1
Reliability[h]												
	Overall Sample	NR	S	S/A[b]	A	S/A/I[e]	S	A	NR[i]	S	NR	S/A[c]
Internal Consistency	K-aged Sample	S	S	S/A[m]	NR	NR	NR	A/I[j]	NR[i]	S	NR	NR
Test-retest reliability	Overall Sample	A	S	A	S	S/A[f]	S	S	I[h] S/I[h]	NR[i]	I	A/I[d]
	K-aged Sample	NR	NR	NR	NR	NR	NR	NR	NR	NR[i]	NR	NR
Interrater reliability[a]	Overall Sample	A	NR[i]	I	A	NR[k]	A	NR[i]	A/I[l]	NR[i]	A/I[g]	NR[k]
	K-aged Sample	NR	NR[i]	NR	NR	NR[k]	NR	NR[i]	NR	NR[i]	NR	NR[k]
Validity evidence												
Content		I	S	I	I	S	I	A	S	S	I	S
Internal structure		I	A	I	I	NR	I	I	NR[i]	A	NR	I
External relations												
Criterion-related		S	A	S	S	S	I	I	I	A	I	I
Group-differences		I	I	I	I	I	I	I	A	NR[i]	NR	A
Consequences		NR	NR	NR	NR	NR	I	NR	I	NR[i]	NR	NR
Response processes		NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Norming data												

Measurement Property	Sample	BASC-3 BESS	BERS2	BRIEF2 Screening Teacher Form	CAS	Conners EC Behavior Short Form	DESSA-Mini	EBS	SSBD	SSBS-2	SSiS Performance Screening Guide	T-CRS 2.1
Recency		S	I	NR	A	I	S	NR	I	I	NA	I
National sampling		S	A	S	NR	NR	NR	A	I	I	NA	I
K-inclusive Sample size		S	S	S	S	S	S	A	S	A	NA	S
Bias evaluation		NR	I	NR	I	A	I	A	NR	A	NR	S

Note. A = adequate, I = inadequate, NR = not reported; S = strong.
BASC-3 BESS = Behavior Assessment Scale for Chidlren-3, Behavioral and Emotional Screening System; BERS2 = Behavioral and Emotional Rating Scale 2; BRIEF2 = Behavior Rating Inventory of Executive Function, Second Edition; CAS = Children Aggression Scale; Conners 3—T(S) = Conners, 3rd Edition–Teacher Short Form; DESSA-Mini = Devereux Student Strengths Assessment; EBS = Emotioanl and Behavioral Screener; EC = early childhood; SSBD = Systematic Screening for Behavior Disorders; SSBS-2 = School Social and Behavior Scales-2; T-CRS 2.1 = Teacher-Child Rating Scale 2.1.
[a]Rated on two categories, adequate and inadequate.
[b]Results from standardization sample demonstrated strong internal correlation coefficient, and results from clinical sample demonstrated adequate internal correlation coefficient.
[c]Results demonstrated adequate internal correlation coefficient for Assertiveness scale, all remaining scales demonstrated strong internal correlation coefficient.
[d]Results demonstrated adequate test-retest coefficient for Task Orientation scale, all remaining scales demonstrated inadequate test-retest coefficient.
[e]Results demonstrated strong for inattention/hyperactivity, adequate for defiant/aggressive behaviors, social functioning/atypical behaviors, and physical symptoms, and inadequate for anxiety and mood and affect.
[f]Results demonstrated strong for inattention/hyperactivity, defiant/aggressive behaviors, social functioning/atypical behaviors, anxiety, and mood and affect, and adequate for physical symptoms.
[g]Results demonstrated adequate for motivation to learn and inadequate for prosocial behavior.
[h]Results demonstrated inadequate for both internalizers and externalizers at Stage 1; results demonstrated strong for the adaptive combined frequency index and inadequate for the maladaptive combined frequency index and the critical events index at Stage 2.
[i]Sample of studies conducted to determine reliability of measure did not include kindergarten-aged students.
[j]Results demonstrated adequate for clinical sample and inadequate for non-clinical sample.
[k]Interrater reliability not reported.
[l]Interrater reliability results at Stage 1 ranged from 0.42 to 0.70; results demonstrated adequate for measures at Stage 2.
[m]Results of K-specific age group separated by gender and demonstrated strong for five to seven year old girls, and adequate for five to seven year old boys.

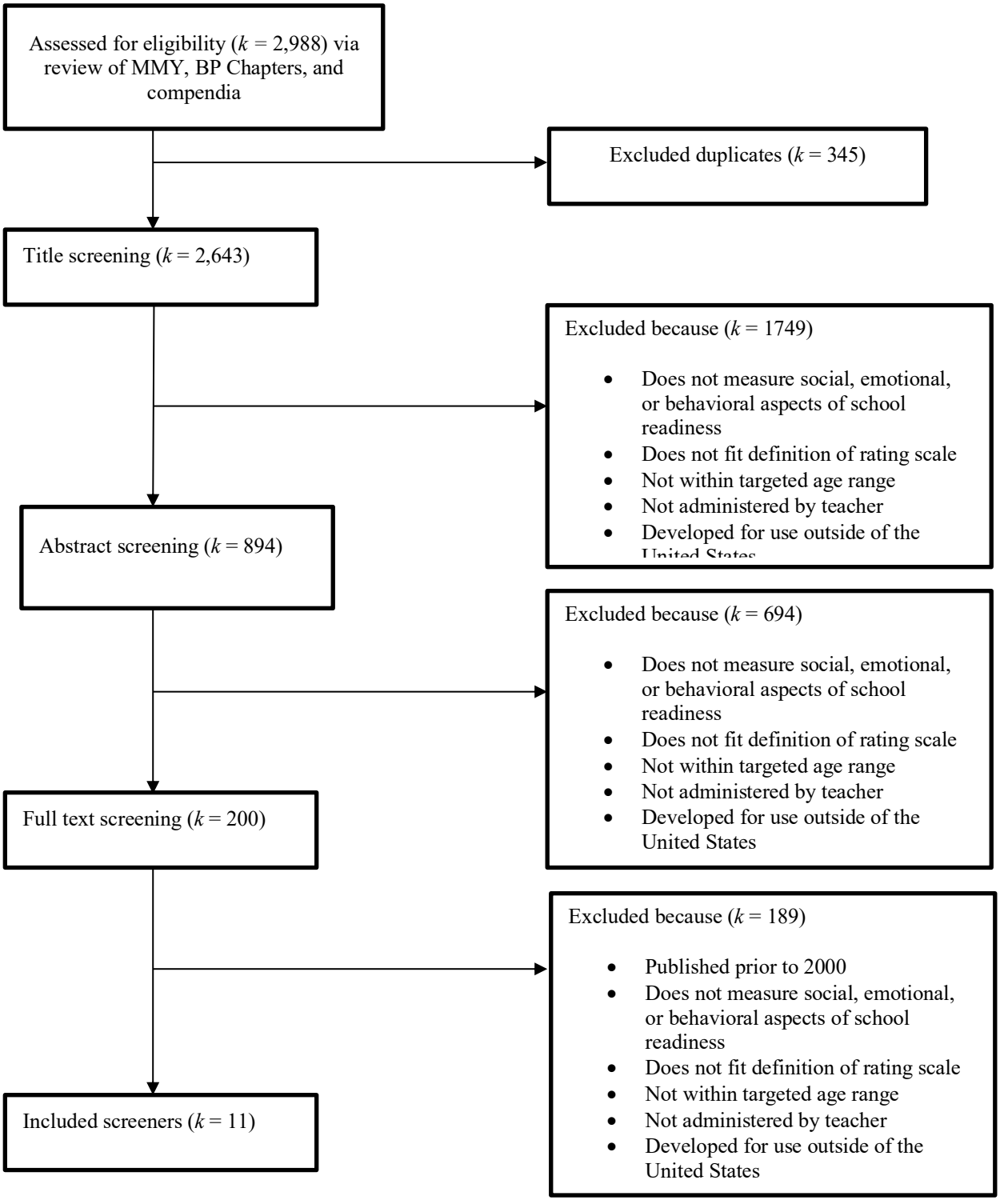

Figure 1. Systematic review process for determining screener eligibility.

Chapter 3

Study 2: A Validation Study of the BASC-3 BESS Teacher Version for a Somali Kindergarten Students

The implementation of universal screening procedures is essential for identifying students who are experiencing social-emotional and behavioral (SEB) difficulties. Best practice strategies include the administration of teacher-reported screening measures to identify students who require additional SEB supports (DiPerna, Bailey, & Anthony, 2014; Glover & Albers, 2007; Ikeda, Neessen, & Witt, 2008). Such practices have been demonstrated to be most effective for identifying students experiencing SEB difficulties above and beyond other methodologies, such as teacher referral (Eklund et al., 2009; Miller et al., 2015). However, careful consideration must be given when selecting screening measures, such that only measures with strong psychometric properties, including analyses of bias evaluation evidence, are utilized. Bias evaluations are particularly important as they provide information regarding the scale's applicability and accuracy for diverse student populations (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014). This is important as kindergarten classrooms across the United States (U.S.) are becoming increasingly diverse (U.S. Department of Education, 2018). Kindergarten is a particularly critical time for developing SEB skills essential for development. Specifically, strong SEB development in kindergarten is positively predictive of future academic performance, whereas poor SEB functioning is predictive of lower language skills and poorer academic development over time (Montes, Lotyczewski, Halterman, & Hightower, 2011; Sabol & Pianta, 2012). Therefore, the

implementation of early universal screening procedures is beneficial to students' educational success.

Unfortunately, bias evaluation procedures are seldom completed at the time of measure publication and are instead often completed via applied research studies following publication of the measure (e.g., Kiperman, Black, McGill, Harrell-Williams, & Kamphaus, 2014). Furthermore, few studies have focused solely on validating SEB screening measures specifically for diverse kindergarten students. The purpose of this study was to expand bias evaluation literature of screening measures for SEB difficulties by examining the validity evidence based on the internal structure of the Behavioral Assessment Scale for Children, Third Edition, Behavioral and Emotional Screening Scale (BASC-3 BESS) Teacher, one of the most frequently administered SEB screeners (Bruhn, Woods-Groves, & Huddle, 2014), specifically for a diverse kindergarten population in a Midwestern state.

Growing Diversity in Schools

The composition of kindergarten students enrolled in United States (U.S.) classrooms is rapidly changing. Over the past few decades, schools across the U.S. experienced an increase in students of various demographic backgrounds, such as students of low-income families, immigrant families, and who identify as English language learners (Digital Promise Global, 2016; Sullivan, Houri, & Sadeh, 2016). In addition, the proportion of students of diverse racial backgrounds has drastically increased in recent years and may increase to more than 50% of the student population by 2027 (National Center of Education Statistics [NCES], 2018, Table 203.60). This is particularly important as students of diverse racial backgrounds are more likely to be

perceived by teachers as experiencing SEB difficulties, thereby increasing the likelihood of future negative outcomes such as increased feelings of stigma, reduced grades, and increased problem behaviors in adulthood (Dever, Raines, Dowdy, & Hostutler, 2016; Fish, 2017; Hosp & Reschly, 2003; Rimm-Kaufman, Pianta, & Cox, 2000; Tenenbaum & Ruch, 2007; Welsh & Little, 2018).

Teacher perceptions of behavior problems in diverse students. Beginning in preschool, teachers' perceptions of behavior problem significantly differ based on student race, and have been more pronounced for Black students (Downer, Goble, Myers, & Pianta, 2016; Gilliam, Maupin, Reyes, Accavitti, & Shic, 2016; Hosp & Reschly, 2003; Rimm-Kaufman et al., 2000; Tenenbaum & Ruch, 2007). Negative perceptions of students early in educational settings have been linked to negative outcomes, such as decreases in academic performance and increases in negative disciplinary actions such as office discipline referrals and in- and out-of-school suspensions (Bradshaw, Mitchell, O'Brennan, & Leaf, 2010; Downer et al., 2016; Downey & Pribesh, 2004; Ferguson, 1998; Skiba, Poloni-Staudinger, Gallini, Simmons, & Feggins-Azziz, 2006; Welsh & Little, 2018). In addition, increased teacher perceptions of problem behaviors are positively related to teacher referrals for special education evaluation for emotional disturbance, particularly for Black students (Dever et al., 2016; Fish, 2017). As such, the implementation of early identification procedures for SEB difficulties that are reliable and valid for all student populations is critical as they may address disproportionality rates in teacher perceptions of problem behaviors and referrals to special education.

Universal screening procedures have been identified as an effective means of accurately identifying students with SEB difficulties and alleviating disproportionalities

among students of racially-diverse background by providing a systematic way of identifying students who are truly in need of SEB supports (Dever et al., 2016; Raines, Dever, Kamphaus, & Roach, 2012). Such practices have been demonstrated as effective specifically for Black students (Dever et al., 2016; Raines et al., 2012). It is important to note, however, the current state of the literature does not consider the substantial cultural variability that exists within the Black student population that can differentially influence student outcomes (Capps & Fix, 2012). Therefore, special attention should be given to this student population specifically in relation to the cultural influence on students' SEB functioning in school.

Growing diversity within Black community. According to the U.S. Census, the Black racial category intends to include all persons who identify as Black or African American who are native to the U.S. and who emigrated from regions such as the Sub-Saharan African and the Afro-Caribbean regions (Rastogi, Johnson, Hoeffel, & Drewery, 2011). This definition fails to consider the cultural heterogeneity that exists in this population (Capps & Fix, 2012). For example, 12% of Black children under the age of 10 living in the U.S. are children of immigrant parents, and more than 80% of Black children of immigrants have a parent from the Sub-Saharan African or the Afro-Caribbean region (Hernandez, 2012). Unsurprisingly, then, the cultural background of Black children is highly diverse in terms of region of origin, language, and other characteristics, leading to great differences within the educational setting, such as teacher ratings of students' kindergarten readiness skills (Crosby & Dunbar, 2012; Hernandez, 2012). Black students' region of origin, then, warrants special attention, particularly when exploring students' early educational experiences. Therefore, this study focused on measuring the

SEB functioning solely for students of Somali descent only to limit the effects of cultural influences from varying Black heritage that may differentially influence SEB functioning among Black students.

As of 2009, Black students of Somali descent comprised approximately nine percent of the Black children of immigrants in the U.S. under the age of 10 (Hernandez, 2012). Of all students of Somali-descent residing in the U.S., three percent resided in Minnesota and constituted the third largest student group enrolled in Minnesota (Darboe, 2003; Hernandez, 2012). Historically, this Midwestern region experienced an influx of Somali immigrants in the early 1990s as a result of the civil war in Somalia (Scuglik, Alarcon, Lapeyre, Williams, & Logan, 2007). Sadly, a large proportion of Somali families have experienced traumatic events prior to immigrating to this region as refugees (Koch, 2007; Scuglik et al., 2007). For this reason, students of Somali descent may be at increased risk of experiencing problem behaviors in schools and may benefit from early SEB supports (Koch, 2007). The early identification of problem behaviors is then crucial for identifying students requiring additional supports as early as possible and can be best attained through the use of universal screening measures (Dever et al., 2016; Raines et al., 2012). However, given the increased risk of negative perceptions of classroom behaviors specifically for Black students, careful consideration must be made in terms of assessment approaches for this population. Bias evaluation procedures have been identified as appropriate methods for examining the psychometric properties of measures across different populations (AERA et al., 2014; Dimitrov, 2010).

Bias Evaluation Evidence

Validation of measures typically begins with evaluations of reliability and validity evidence with nationally-representative norming samples (AERA et al., 2014). Subsequently, test developers are responsible for completing bias evaluations of the measures to demonstrate its equivalence for varying populations (AERA et al., 2014). Bias can be evaluated by exploring the overall structural coefficients of the measure to ensure the structure holds across populations, referred to as covariance invariance (AERA et al., 2014; Dimitrov, 2010; Sass & Schmitt, 2013). Bias can also be explored by evaluating the properties of individual items on the measure across populations (AERA et al., 2014; Sass & Schmitt, 2013). Bias evaluations typically begin by exploring the structure of the entire measure and end with analyses of the individual items as described below.

Bias evaluation procedures often begin with an assessment of the metric invariance to confirm that the observed variables, or individual items on a measure, load onto the unobserved, latent structures in similar fashion across diverse populations (Dimitrov, 2010; Rios & Wells, 2006). Second, scalar invariance is conducted to examine the extent to which thresholds of the observed variables are similar across populations (Dimitrov, 2010; Rios & Wells, 2006). Third, an evaluation of strict factorial invariance is conducted to ensure equal residuals of observed variables across different populations (Dimitrov, 2010; Rios & Wells, 2006). Finally, item-based analyses are conducted through differential item analyses to explore the individual parameters of each item on the measure across different populations (Dimitrov, 2010; Rios & Wells, 2006).

A critical starting point for the evaluation of bias for a newly developed measure is through an examination of the factor structure. Therefore, this study will examine the

factor structure of the BASC-3 BESS Teacher, specifically for students of Somali descent residing in a Midwestern state. This study will explore the metric invariance of the BASC-3 BESS Teacher for students from Somali descent compared to the overall norming sample through an analysis and comparison of the overall fit of the factor structure and the factor loadings of observed variables (Kamphaus & Reynolds, 2015).

BASC-3 BESS

The BASC-3 BESS Teacher (hereby referred to as *BESS-3 Teacher)* is a rating scale used to assess students' risk for SEB functioning for students from kindergarten through 12th grade (Kamphaus & Reynolds, 2015), and was recently reported as the most frequently used universal screening measure in schools across the U.S. (Bruhn et al., 2014). The BESS-3 Teacher is a 20-item measure that provides a single composite T-score, the Behavioral and Emotional Risk Index (BERI), and three additional subindex raw scores that indicate risk for specific SEB domains. The internalizing risk index (IRI) includes items related to behaviors that are internal in nature and are not disruptive. The externalizing risk index (ERI) consists of items associated with externalizing behaviors that are disruptive. Finally, items representing the adaptive skills risk index (ARI) include adaptive skills essential for functioning at school, with peers, at home, and in the community. The BERI provides cut-off scores for identifying students with SEB difficulties, whereas the sub-index scores provide guidance on the specific SEB domain for which students require additional supports.

The final items included in the BESS-3 Teacher were obtained from the BASC-3, a comprehensive measure of problem and adaptive behaviors used to provide a comprehensive description of students' behaviors (Reynolds & Kamphaus, 2015). Items

with the highest factor loadings for each sub-index of the BASC-3 were included in the BESS-3 Teacher. Once the items were identified, a confirmatory factor analysis (CFA) was conducted to evaluate the overall fit of a second-order factor model to ensure all items presented adequate factor loadings for each of the three first-order latent structures. A total of 17 items mapped onto one of the three first-order latent structures. The remaining three items mapped directly onto the secondary factor, the BERI score. All loadings were identified to be 0.50 or above (Kamphaus & Reynolds, 2015).

The BESS-3 Teacher has been significantly modified from its previous version. First, the BESS-2 Teacher was a 27-item measure and was revised to a 20-item measure as the BESS-3 Teacher. In addition, 12 items were removed from the measure, and five new items were included. One item was slightly reworded. Finally, the scoring procedures were also revised between the last and most recent versions of the BESS-3 Teacher. Results of the BESS-2 Teacher provided only one composite score, while results of several factor analyses studies conducted by researchers in the field demonstrated the existence of first-order latent structures similar to those identified in the BASC-2 (Dever, Mays, Kamphaus, & Dowdy, 2012; Dowdy et al., 2011). Therefore, the current BESS-3 Teacher was modified to provide one composite score and three subindex scores similar to the BASC-3 (Kamphaus & Reynolds, 2015; Reynolds & Kamphaus, 2015). Given the extent of the modifications made on the current version of the BESS-3 Teacher, further examination of its factor structure is warranted. Furthermore, the factor structure of the BESS-3 Teacher has not been examined for students of racially-diverse backgrounds. Therefore, this study will explore the factor structure of the BESS-3 Teacher specifically for students of Somali descent.

Current Study

As classrooms across the U.S. continue to increase in diversity (NCES, 2018, Table 203.60), school personnel need to identify school practices that are equitable across all students, particularly for the identification of students exhibiting SEB difficulties. Current practices that involve teacher referrals may be linked to biased perceptions of problem behaviors, which may result in detrimental consequences, particularly for Black students (Bradshaw et al., 2010; Downer et al., 2016; Downey & Pribesh, 2004; Ferguson, 1998; Gilliam et al., 2016; Skiba et al., 2006; Welsh & Little, 2018). As such, school personnel may incorporate universal screening procedures beginning in kindergarten that would more equitably identify students at-risk for SEB difficulties (Dever et al., 2016; Raines et al., 2012). However, such screening measures must first demonstrate measurement invariance across all students within their student population (AERA et al., 2014). For this reason, researchers and test developers will need to more carefully attend to the ways in which students' diverse backgrounds influence test results by more thoroughly conducting bias evaluation analyses (AERA et al., 2014). Unfortunately, bias evaluation procedures are seldom completed at the time of measure publication and are often completed via applied research studies following publication of the measure (e.g., Kiperman et al., 2014). Furthermore, few studies have focused solely on validating SEB screening measures specifically for diverse kindergarten students. Therefore, the purpose of this study was to examine the overall factor structure of the BESS Teacher, specifically with kindergarten students of Somali descent.

To evaluate the internal structure of the BESS-3, both the dimensionality and the reliability of the BESS-3 Teacher were evaluated (Rios & Wells, 2014). The

dimensionality of the measure were explored through the use of confirmatory factor analysis (CFA), which provides empirical evidence for the factor structures identified with the norming population of the BESS-3 Teacher for students of Somali descent. This type of analysis allows for a better understanding of the dimensions of the constructs being measured in this population. Lastly, this type of analysis provides evidence that the measurement of the latent variables is psychometrically sound (Byrne, 2012; Rios & Wells, 2014). Next, this study examined the interrelationships of the items of the BESS-3 Teacher for each subindex as a form of internal consistency. This provides evidence the items within each subindex are consistently measuring the same constructs. Therefore, this study aimed to answer the following questions:

1. Does the higher-order single-factor model of the BESS-3 Teacher identified by the test developers (Kamphaus & Reynolds, 2015) result in a well-fitted model when applied to a Somali kindergarten population as determined by the goodness of fit indices and individual factor loadings?
2. Do the BESS-3 Teacher items identified within each latent factor correlate with each other for the Somali student population as demonstrated by a correlation of 0.80 or greater (Alfonso & Flanagan, 2009; Bracken, 1987), indicating adequate internal consistency?
3. Do the subindex scores correlate with the BERI score with the Somali student population with correlations above 0.70 as they do with the norming population (Kamphaus & Reynolds, 2015)?

Methods

Participants

Data collection was conducted at multiple charter schools in a Midwestern state during the 2018-2019 school year. A total of five charter schools initially agreed to participate, however, one charter school withdrew from the study prior to the start of data collection. To be included in this study, students needed to be in kindergarten, be 5 years of age or older, and be of Somali descent. Twelve teachers initially completed the consent process for this study; however, three teachers (25%) withdrew from the study prior to the completion of any student rating scales.

Overall, a total of 161 rating scales were completed and submitted by nine teachers. One rating scale was excluded as it could not be accurately scored due to missing 50% of the individual items of the BESS-3 Teacher. As a result, a total of 160 ratings scales were included in this study. The 160 rating scales were completed for students who were 50% female with a mean age of 5.88 years.

Table 3 provides teacher demographic information. Overall, most teachers who participated in the study were female, White, and non-Hispanic. Furthermore, teachers reported on class-wide student demographic information presented as percentages. Table 4 provides students' demographic for each school. Across all classrooms, students' race was reported to be Black, Somali, or Latino/a. Further, all students were reported to meet eligibility for free or reduced lunch. Lastly, approximately 11% of students received an office discipline referral less than 1 month prior to data collection.

Materials and Procedures

BASC-3 BESS Teacher rating form. The BESS-3 Teacher is a 20-item universal screening form that takes less than five minutes per student to complete. Teachers require little to no training prior to completing the BESS-3 Teacher. However,

teachers need to have daily contact with students for approximately a month prior to completing the form. Results of the BESS-3 Teacher provide the BERI, a T-score representing overall emotional and/or behavioral risk for the student. In addition, the BESS-3 Teacher also provides subindex raw scores for the IRI, ERI, and ARI that may be used to identify the risk level of the student for each subindex.

Reliability and validity evidence were measured using a norming sample representative of the general population. The norming sample included approximately 1,600 students recruited from 44 states across the U.S. from April 2013 through November 2014. Approximately 150 students in the norming sample were either 6 or 7 years of age at the time of data collection, including kindergarten students, with an equal number of male and female students. Demographic characteristics of the normative sample were comparable to the characteristics of the U.S. population as demonstrated by the data collected by the 2013 U.S. Census Bureau American Community Sample. Demographic characteristics included student race or ethnicity (African American, Asian, Hispanic, Other, and White), mother's education (less than a high school diploma, high school graduate, some college or technical school, and four years of college or more), and the family's region of residence (Northeast, Midwest, South, or West). Specifically, approximately 51% of students 6 or 7 years of age identified as White, 25% as Hispanic, 14% African American, and 5% each of students identified as Asian or Other.

Internal consistency evidence measured specifically for students between the ages of six and seven ranged 0.82 to 0.95 as measured by coefficient alpha. In addition, test-retest reliability evidence ranged from 0.82 to 0.89, and interrater reliability ranged from 0.52 to 0.67. Validity based on the internal structure was demonstrated through index and

subindex intercorrelations. Correlations between the BERI and subindex scores ranged from 0.68 to 0.88. The ERI demonstrated to have the highest correlations with other subindex scores, while the ARI demonstrated the lowest correlations with other subindex scores. Validity based on relations with external variables was measured via correlations with other emotional and behavioral measures. Specifically, the correlations between the BESS-3 Teacher index score and the total scores of the BASC-3 Teacher (r = 0.92), the Achenbach System of Empirically Based Assessment Teacher Report form (r = 0.78), the Conners 3rd Edition Teacher Form (r =0.66), and the Autism Spectrum Rating Scale (r = 0.57) were provided. In addition, correlations between the index and subindex scores of the BESS-3 Teacher and the index score BASC-2 BESS ranged from 0.65 to 0.97, indicating a strong relationship between the two versions of the BESS. Lastly, predictions of the BASC-3 Teacher total score based on the BESS-3 Teacher results were calculated and demonstrated acceptable levels of sensitivity (0.73) and specificity (0.95). The positive predictive power and negative predictive power for base rates of 10% to 50% ranged from 0.26-0.76 and 0.95 to 0.99, respectively.

Procedures. Kindergarten teachers provided consent for study participation prior to enrolling in the study. Although parent consent was deemed unnecessary for this study as no private information were collected and the procedures were consistent with the schools' typical functioning, one school opted to send home a parent notification letter to inform parents of the study. Once consented, teachers completed the BESS-3 Teacher for each student who met eligibility criteria for the study. Specifically, teachers were instructed to complete the rating form with all Somali students who were at least 5 years of age and with whom they had regular contact with for at least one month. All teachers

completed the BESS-3 Teacher electronically via Q-global. Teachers provided each eligible student's age and gender prior to the completion of the rating forms. Each student was then assigned a unique ID number to preserve student anonymity. Teachers maintained the master list that linked students' ID numbers to student names and were instructed to shred this information upon completion of study procedures. Teachers received a $50 gift card once they completed the rating forms as compensation for their time.

Analytic Plan

Two statistical packages were used for this study. The Statistical Package for the Social Sciences (SPSS) version 25.0 was used to calculate available descriptive statistics, and to conduct correlational analyses described below. CFA models were analyzed using M*plus* Version 8.2 (Muhtén & Muhtén, 2017). The following analyses were completed to answer the study's research questions.

Preliminary analyses. Preliminary analyses were conducted to provide study demographic information. First, descriptive statistics were calculated, and included student percentage of gender and students' mean age. Second, necessary statistical assumptions required for the CFA were evaluated. These include the identification of any outliers, and an examination of the distribution of the variables. Specifically, the data were examined to determine if they followed a normal distribution as identified by visual analyses of histograms and calculations of skewness and kurtosis (Brown, 2015; Dimitrov, 2010). In addition, scatterplots were evaluated to assess the data's linearity and homoscedasticity (Brown, 2015; Dimitrov, 2010). It is important to note, however, that the estimator identified for use for the CFA, the weighted least squares using diagonal

weight matrix (WLSMV) estimation method, accounted for both the categorical nature of the variables of the BESS-3 Teacher and the non-normally distributed data in the case that the study data were determined to be non-normally distributed (Brown, 2015; Byrne, 2012; Dimitrov, 2010).

Confirmatory factor analysis. CFA of the BESS-3 Teacher was conducted using M*plus* Version 8.2 (Muhtén & Muhtén, 2017). The CFA model was conducted following the model used by Kamphaus and Reynolds (2015) at the time of development of the BESS-3 Teacher. Specifically, a second-order factor model with three first-order factors was tested. This model hypothesized the presence of a BERI higher-order factor, and three first-order factors: IRI, ERI, and ARI. A total of 17 items were mapped onto one of the three first-order factors. The remaining three items were mapped directly onto the second-order factors as they did not fit within any of the identified first-order factors (Kamphaus & Reynolds, 2015). WLSMV was used to estimate the model parameters. The WLSMV approach accounted for the categorical outcome of the BESS-3 Teacher and the skewed or non-normal distribution of the data, and was appropriate for use with smaller sample sizes (Brown, 2015; Bryne, 2012). Lastly, coefficient omega was calculated as a measure of internal consistency for each of the model factors (Kelley & Pornprasertmanit, 2016; Rios & Wells, 2014). Coefficient omega was included as part of the analyses as it has been demonstrated to be a more accurate and reliable measurement of internal consistency as compared to Cronbach's alpha (Kelley & Pornprasertmanit, 2016; Viladrich, Angulo-Brunet, & Doval, 2017; Sijtsma, 2009).

Using results from the CFA, model fit indices were then compared to those obtained through the second-order factor model employed by Kamphaus and Reynolds

(2015). According to the technical manual, the CFA model indicated an adequate fit as presented by a comparative fit index (CFI) of 0.91 and root mean square error of approximation (RMSEA) value that fell in the 0.06 to 0.1 range. It is important to note good model fit indices as described by the literature demonstrate RMSEA values close to 0.06 or below, standardized root mean square residual (SRMR) values close to 0.08 or below, and CFI and Tucker-Lewis index (TLI) values greater 0.95 (Hu & Bentler, 1990). However, model fit indices with RMSEA values less than 0.08 and SRMR values less than 0.1 may suggest adequate fit (Brown, 2015; Browne & Cudeck, 1993; MacCallum et al., 1996). Factor loadings for individual items presented by Kamphaus & Reynolds (2015) were demonstrated to be greater than 0.50.

Correlations. The split-half method was used to assess the internal consistency of the BERI scores (Kamphaus & Reynolds, 2015). Each half contained a balance of items that represented similar behavioral problems, which were correlated. The Spearman-Brown formula was then applied to the correlation to obtain a final measure of the internal consistency of the BERI. Additionally, the coefficient alpha was measured for each subindex score in order to assess and compare the consistency of each subindex to those reported in the BESS-3 manual (Kamphaus & Reynolds, 2015). Lastly, the internal structure of the BESS-3 Teacher was measured through examination of the interrelationships of the subdomain scores with the BERI total score through Spearman correlations (AERA et al., 2014; Kamphaus & Reynolds, 2015).

Results

Preliminary Analyses Findings

Assumption testing. First, study data were explored to identify degree of skewness and kurtosis (see Table 6). Results indicated 15 of the study items appear to be non-normally distributed, demonstrating marked skewness. For this reason, WLSMV methods of estimation were used to account for the non-normality of the data (Brown, 2015; Byrne, 2011). This estimation method further accounted for the categorical nature of the variables as well as the small sample size (Byrne, 2011).

Confirmatory Factor Analysis Findings

This study aimed to replicate the second-order CFA model employed by Kamphaus & Reynolds (2015) to determine if this model fits well with a Somali student population. First, an evaluation of the model fit was conducted to determine if this model fit the data well. Individual item and latent variables factor loadings were then calculated and compared to the results provided by Kamphaus and Reynolds (2015).

Results of this study are presented in Tables 5 and 6. First, results of the chi-square test of model fit (χ^2 (167, N = 160) = 400.84, $p < 0.001$) demonstrated a poor fit of the data. However, due to the sensitive-nature of chi-square model fit test to non-normally distributed data, large sample sizes, and the very stringent null hypothesis that the observed matrices are equal to those predicted, the results of the chi-square model fit test may falsely indicate poor fit (Browne, 2015; Byrne, 2011). For this reason, multiple tests of model fit were completed. CFI and TLI values were found to be greater than 0.95 (Brown, 2015; Hu & Bentler, 1995). In addition, SRMR was demonstrated to be less than 0.09 (Brown, 2015; Hu & Bentler, 1995). Although the RMSEA value falls above the

cut-off provided by the literature (Brown, 2015, Browne & Cudeck, 1993; MacCallum, Browne, & Sugawara, 1996), the RMSEA identified by this study was found to be within the range provided by Kamphaus & Reynolds (2015). Overall, the results of the CFA model implemented by this study suggest adequate model fit and may be interpreted and compared to those obtained from the normative sample of the BESS-3 Teacher (Kamphaus & Reynolds, 2015).

Results of the second-order factor model demonstrated factor loadings greater than 0.50 for individual items and the first-order latent structures. Factor loadings ranged from 0.81 to 0.98 for the ERI, 0.62 to 0.91 for the IRI, and 0.84 to 0.91 for the ARI. Such factor loadings indicated a strong positive predictive relationship between each first-order latent variable and the individual items comprised within each latent variable. Further, factor loadings between the BERI and the three individual items ranged from 0.84 to 0.95. Lastly, the factor loadings associated between the BERI and ERI, IRI, and ARI were identified to be 0.76, 0.57, and -0.86, respectively. These results demonstrated positive predictive relationships between the BERI and the individual items and the IRI and ERI, and a negative predictive relationship between the BERI and the ARI. Lastly, the internal consistency of the BERI, ERI, IRI, and ARI as measured by coefficient omega were 0.88, 0.96, 0.89, and 0.94, respectively, demonstrating adequate to strong internal consistency for all factors (Alfonso & Flanagan, 2009; Bracken, 1987).

Correlations Findings

Correlational results of this study are presented in Tables 7 and 8. Results of this study demonstrated a lower internal consistency measurement for the BERI ($r = 0.84$) than reported by Kamphaus & Reynolds (2015; $r = 0.95$), although psychometrically

acceptable (Alfonso & Flanagan, 2009; Bracken, 1987; Floyd et al., 2015). The internal consistency of the subindices ranged from 0.83 to 0.90, falling within the range reported by Kamphaus & Reynolds (2015). Intercorrelations between the BERI and subindices scores measured by this study ranged from 0.66 to 0.87 and are similar to those reported by Kamphaus & Reynolds (2015). Further, the ERI demonstrated the highest correlation with the other subindex scores, consistent with findings reported by Kamphaus & Reynolds (2015). However, this study identified the IRI, not the ARI (Kamphaus & Reynolds, 2015), as having the lowest correlations with other subindex scores.

Discussion

The systematic use of teacher rating scales as part of universal screening procedures have been demonstrated to reduce disproportionate rates of referrals for SEB difficulties, particularly for diverse students (Dever et al., 2016; Raines et al., 2012). It is crucial, however, that the identified teacher rating scales demonstrate measurement invariance for diverse students. As such, this study sought to provide preliminary metric invariance evidence to support the use of the BESS-3 Teacher in a Somali kindergarten population, an essential first step in bias evaluation procedures. Overall, this study demonstrated a Somali kindergarten sample produced psychometric characteristics similar to those obtained with the normative sample of the BESS-3 Teacher, providing preliminary support for its use with a Somali kindergarten student population.

First, this study assessed the internal structure of the BESS-3 Teacher and compared the results to those reported in the BESS-3 manual. Specifically, the results of this study demonstrated adequate model fit indices when following the guidelines set by the literature (Brown, 2015). However, due to limited reporting in the manual, it is

unclear if the model fit indices obtained by this study are similar to those obtained by the normative sample. Specifically, the BESS-3 manual reported only one absolute (RMSEA) and one incremental (CFI) fit index. However, the inclusion of additional model fit indices would provide stronger evidence of the internal validity of the rating scale (Hu & Bentler, 1995). Furthermore, the BESS-3 manual did not provide the model fit indices specifically for the Teacher version of the scale. Instead, authors reported the range of values for the model fit indices across all versions of the rating scale (Kamphaus & Reynolds, 2015). This is particularly problematic as it is unclear which of the model fit indices are associated with the BESS-3 Teacher, limiting possible interpretations of model fit. Although model fit could not be thoroughly compared between the two samples, the reporting in the manual was sufficiently detailed to permit an analysis of the individual factor loadings between samples.

To this end, the CFA results of this study demonstrated that the individual factor loadings were comparable to those reported in the BESS-3 manual (Kamphaus & Reynolds, 2015). This indicates that the items loaded onto each factor appropriately and consistently represent each construct across both the normative and Somali kindergarten population (Brown, 2015). In addition, while the BESS-3 manual failed to report the factor loadings of the first-order factors mapped onto the second-order factor, this study demonstrated that the first-order factor structures adequately map onto the second-order factor structure, providing further preliminary support for the use of a second-order model. These findings provide preliminary evidence supporting the metric invariance of the BASC-3 BESS Teacher specifically for a Somali kindergarten population by

demonstrating that the individual items load onto the latent structure in a similar fashion across the different populations (Rios & Wells, 2014).

It is important to note this study specifically selected the WLSMV estimation method to extract factor loadings of the BESS-3 items. This estimation method was proactively chosen to account for both the categorical nature of the items as well as the violation of the normality assumption of the study data. Due to the limited reporting in the BESS-3 manual specifically in relation to the CFA conducted, it is possible that the BESS-3 authors opted to use other estimation methods, such as maximum likelihood. Such procedures may impact the results and may make comparisons between findings of this study and those identified by the BESS-3 Teacher manual difficult, highlighting the importance of thoroughly reporting procedures used to evaluate psychometric properties of rating scales (AERA, APA, & NCME, 2014; Brown, 2015; Byrne, 2012).

Next, this study sought to explore the internal consistency and reliability of the BESS-3 Teacher in a Somali kindergarten population. This study measured the internal consistency and subindices of the BERI by replicating the procedures completed in the BESS-3 manual, and by calculating the omega coefficient of the factors identified in the CFA model. These procedures expanded on the results provided in the BESS-3 manual and provided additional reliability evidence of the BESS-3 Teacher. Compared to coefficient alpha, coefficient omega has been demonstrated to provide a more accurate measure of internal consistency as it accounts for variability in the factor loadings within each construct (Kelley & Pornprasertmanit, 2016; Viladrich et al., 2017). As a result, the reliability of the subindices as measured by coefficients alpha and omega was

demonstrated to be adequate and comparable to those obtained with the norming sample (Kamphaus & Reynolds, 2015).

However, there were differences noted between this study and the original norming sample. The internal consistency of the BERI, as measured by both the Spearman-Brown correlation and the coefficient omega, was deemed to be lower in the Somali kindergarten sample compared to the norming sample reported in the BESS-3 manual (Kamphaus & Reynolds, 2015). A lower internal consistency metric indicates some inconsistencies exist in the responses provided across all of the items of the BESS-3 Teacher, specifically in the study population. However, the reliability of the BERI was determined to be adequate based on the guidelines in extant literature (Alfonso & Flanagan, 2009; Bracken, 1987), and only appears to be low when compared to the reliability of the BERI as demonstrated by the norming sample (Kamphaus & Reynolds, 2015). Therefore, it can be concluded the internal consistency of the BERI can be accurately interpreted when measured within a kindergarten Somali population.

Although scarce, bias evaluation studies are crucial for demonstrating adequate reliability and validity of rating scales for diverse populations. Given the recent development of the BESS-3 Teacher, evidence of measurement invariance of this rating scale in diverse populations is limited. Consequently, many of the items of the BESS-3 Teacher were retained from its predecessor, the BASC-2 BESS Teacher, signifying that previous bias evaluation studies of the BESS-2 Teacher may provide preliminary evidence for the use of the BESS-3 Teacher in diverse populations. Although several studies explored measurement invariance of both parent and student versions of the BASC-2 BESS (e.g., Dever, et al., 2016; Kiperman, et al., 2014), support for the use of

the teacher version of the BASC-2 BESS specifically has been provided only for Hispanic youth (Wiesner & Schanding, 2013) and youth residing in rural schools (King, Reschly, & Appleton, 2012). Such limitations are problematic both in the literature and in practice, particularly for the BESS Teacher, as it was identified as the most widely used screening measure in the United States (Bruhn et al., 2014). Without evidence of measurement invariance, screening results may disproportionally rate diverse students as being at high-risk for SEB difficulties, which may lead to inaccurate or inappropriate identification procedures. Therefore, future research must intentionally include representation of diverse students in bias evaluation studies of rating scales.

Importantly, when planning future bias evaluation studies, researchers should consider greater precision of reporting and representation of specific racial and ethnic groups. Racial categories often used in the United States are over-inclusive and fail to consider the cultural heterogeneity that exists in each racial category, which may lead to negative implications in test use and interpretation (Braun et al., 2007). In a school setting, failure to consider cultural differences may negatively influence results of rating scales, ultimately leading to disproportionate identification of students for interventions or special education referrals (Hernandez, 2012). This study provided a first step in the consideration of cultural heterogeneity that exists specifically in the Black community by exploring the psychometric properties of the BESS-3 Teacher solely for a Somali student population. This was particularly important as behavioral differences based on cultural factors have been identified in this community (Hernandez, 2012). Future studies may then continue to expand on the literature by further exploring the implications of cultural

heterogeneity in each racial category on the psychometric properties of behavioral rating scales.

Limitations and Directions for Future Research

The findings and interpretations of this study are tempered by its limitations. First, this study employed a small sample of teachers from a limited number of schools, which resulted in a small study sample. Small study samples are at increased risk for not meeting normality assumptions, which limit study data interpretations. Further, small study samples lead to limitations in generalizability. A larger sample would be more likely to meet assumptions of normality, and would likely lead to more generalizable results. Future studies should expand on this study by including larger samples of students recruited from various schools across the nation.

In addition, this study demonstrated that a higher-order factor model appears appropriate when compared to the findings reported in the BESS-3 manual. However, the model fit indices may be better improved with different models, such as bifactor models, or the removal of the three additional items that map directly onto the second-order structure. Future studies may consider conducting exploratory factor analyses particularly with diverse student populations to ensure the higher-order factor model is the most appropriate within diverse student populations.

Lastly, this study provided preliminary evidence for the use of the BESS-3 Teacher solely for a Somali student population. Notably, this study did not include students of other diverse racial and cultural backgrounds, failing to compare the findings of this study to students of other diverse backgrounds. Therefore, future bias evaluation studies must be conducted with various diverse student populations to provide evidence

for the use of the BESS-3 Teacher for those specific populations, and identify any cultural differences that may exists between ratings of students from diverse backgrounds. Such procedures may then provide stronger evidence for measurement invariance across diverse populations (Brown, 2015).

This study also identified significant gaps in the literature that must be addressed. Specifically, this study focused solely on kindergarten students, which results in a study sample mean age of 5.88. However, while the manual states that the BESS-3 Teacher scale can be administered as early as fall of kindergarten, the age range provided in the manual begins at six years of age. This may be problematic as most students enroll in kindergarten at five years of age (NCES, 2013). Further, teacher ratings of SEB functioning tend to demonstrate lower rates of SEB problems for older kindergarten students (Datar & Gottfried, 2015), which may significantly impact norming results of a rating scale. Therefore, future studies must explore the factor structure of five-year-old students and compare it to the findings reported in the BESS-3 manual (Kamphaus & Reynolds, 2015).

Lastly, this study focused solely on examining the internal structure of the BESS-3 Teacher through CFA and correlational analyses. However, no other psychometric properties were explored for this Somali kindergarten student sample. It is essential that all psychometric properties are explored to ensure the rating scales are valid and reliable for all student populations (AERA et al., 2014). Further, this study only provided one source of evidence for the validity based on internal structure, while the literature states that a minimum of three sources of evidence must be presented (Alfonso & Flanagan, 2009; Bracken, 1987; Floyd et al., 2015, Halle & Darling-Churchill, 2016). Therefore,

future studies must expand on this study by assessing the remaining psychometric properties via several sources of evidence to truly demonstrate the validity and reliability of the BESS-3 Teacher for a Somali kindergarten population.

Implications for Practice

Relying solely on teacher perception for behavioral problems particularly for diverse students has many limitations, including increases in negative disciplinary actions and decreases in academic performance (Bradshaw et al., 2010; Downer et al., 2016; Downey & Pribesh, 2004; Ferguson, 1998; Skiba et al., 2006; Welsh & Little, 2018). The implementation of universal screening procedures, such as those that include teacher report rating scales like the BESS-3, has been shown to be an effective strategy for reducing the disproportionate identification of diverse students for behavioral concerns (Dever et al., 2016; Raines, et al., 2012). Although the psychometric evidence of various ratings scales demonstrates measurement invariance for diverse students, many rating scales fail to consider the cultural heterogeneity of Black students when demonstrating validity and reliability evidence. Therefore, school administrators must be proactive in identifying rating scales that demonstrate validity and reliability evidence for their specific student population. This study demonstrates preliminary evidence that the BESS-3 Teacher rating may be a valid and reliable rating scale for use in a kindergarten Somali population. School administrators who are interested in implementing universal screening procedures for predominantly Somali kindergarten classrooms may preliminarily consider the use the BESS-3 Teacher rating scale while remaining cognizant that its psychometric properties must continue to be evaluated for this student population.

In addition, when implementing the BESS-3 Teacher as part of universal screening procedures in classrooms of predominantly Somali students, school personnel should consider focusing on the results of the subindex scores when identifying students at-risk for SEB difficulties as opposed to the use of the BERI scores, given the adequate reliability measures identified in this study. This is important as the use of the subindices scores provided by the BESS-3 Teacher may provide more specific guidance for the identification of intervention plans for students as opposed to the BERI scores. For example, while the BERI may identify students with SEB concerns, the subindices provide further information by identifying internalizing, externalizing, or adaptive risk specifically. In addition, this study identified a negative relationship between the BERI and ARI, which may make the interpretation of the BERI independent of the subindices difficult. Therefore, school personnel may obtain additional information on the SEB concerns of students by analyzing and interpreting subindex scores.

Conclusion

Results of this study provided preliminary evidence for the use of the BESS-3 Teacher rating scale in a Somali kindergarten population. Specifically, the results of this study demonstrated adequate-to-strong internal consistency and validity based on the internal structure for the BESS-3 Teacher. Therefore, this study provided evidence to support the use of this scale with Somali kindergarten students and classrooms. Future research efforts should continue exploring the remaining psychometric properties of the BESS-3 Teacher in a Somali student sample to provide further evidence for its use with this student population.

Table 3
Teachers' Demographic Characteristics

Characteristics	*n*	*%*
Female	8	88.90
Race		
White	8	88.90
Black	1	11.10
Ethnicity – non-Hispanic	9	100
Total Years in Teaching Profession		
Less than 1 year	1	11.10
1-5 years	6	66.70
6-10 years	1	11.10
More than 20 years	1	11.10
Total Years in Current Position		
Less than 1 year	4	44.40
1-5 years	4	44.40
6-10 years	1	11.10
Highest Earned Degree		
Bachelor's degree	6	66.70
Master's degree	3	33.30
Area of Certification—General Education	9	100

Table 4

Percentage and Standard Deviations of Class-wide Student Demographic Characteristics per School

Demographic Characteristics	School 1	School 2	School 3	School 4
Student Race				
Somali	90.33 (11.93)	97.00 (2.65)	78	97.00 (4.24)
Black	5.33 (5.51)	3.00 (2.65)	21	---
Latino/a	5.33 (5.51)	---	---	---
Other	---	---	---	3.00 (4.24)
Student Services				
FRL	100.00 (0.00)	52.67 (50.21)	100	100.00 (0.00)
EL services	74.00 (45.03)	---	73	12.75 (0.36)
SPED	7.00 (3.46)	0.02 (0.03)	0.05	3.00 (4.24)
ODR	10.67 (11.02)	23.67 (22.59)	10	12.50 (17.68)

Note: No standard deviations are available for school 3 as only one classroom participated in the study. EL = English language; FRL = free/reduced lunch; ODR = office discipline referral; SPED = special education

Table 5
Fit indices of the Second-Order Factor Model

Model	χ^2 Model Fit	RMSEA	SRMR	CFI	TLI
Second-order model	χ^2 (167, N = 160) = 400.84, p < 0.001	0.09	0.09	0.97	0.97

Table 6
Standardized Factor Loadings of the Second-Order Factor Model

Item	Factor Loading	*M*	*SD*	Skewness	Kurtosis
First-Order Factor Structures					
Externalizing Risk					
Poor self-control	0.90	1.88	0.92	0.80	-0.24
Gets into trouble	0.98	1.96	0.79	0.61	0.10
Annoys others	0.81	1.66	0.79	1.00	0.27
Defies teachers	0.87	1.58	0.74	1.23	1.25
Disrupts others	0.92	1.82	0.82	0.77	0.04
Disobeys	0.89	1.79	0.76	0.73	0.21
Internalizing Risk					
Worries	0.69	1.74	0.64	0.43	0.06
Sad	0.73	1.69	0.59	0.19	-0.60
Changes mood	0.91	1.55	0.70	1.11	0.78
Easily upset	0.89	1.88	0.82	0.57	-0.40
Easily stressed	0.69	1.60	0.72	1.08	0.88
Appears tense	0.62	1.43	0.92	1.17	0.31
Adaptive Risk					
Well organized	0.86	2.61	1.00	-0.03	-1.07
Motivated to succeed	0.86	2.83	0.92	-0.23	-0.89
Helps others	0.84	2.36	0.82	0.28	-0.38
Gets others to work	0.91	2.21	0.83	0.33	-0.39
Good study habits	0.91	2.49	0.95	0.06	-0.91
Second-Order Factor Structure					
Behavioral and Emotional Risk Index					
Short attention span	0.95	2.08	0.97	0.59	-0.61
Trouble concentrating	0.94	2.09	0.97	0.46	-0.82
Incorrect assignments due to not following directions	0.84	1.82	0.88	0.87	-0.04
Internalizing Risk	0.76	3.89	3.02	0.89	1.20

Item	Factor Loading	*M*	*SD*	Skewness	Kurtosis
Externalizing Risk	0.57	4.68	4.14	0.96	0.75
Adaptive Risk	-0.86	7.49	3.85	0.03	-0.76

Table 7
Coefficient Alpha Measured for Each Sub-index Score

	Coefficient Alpha
Adaptive Risk	0.90
Externalizing Risk	0.93
Internalizing Risk	0.83

Table 8
Intercorrelations of Sub-indices and BERI Scores

	BERI Score	Adaptive Risk Score	Externalizing Risk Score
BERI Score			
Adaptive Risk Score	-0.87†		
Externalizing Risk Score	0.84†	-0.60†	
Internalizing Risk Score	0.66†	-0.37†	0.48†

*$p < 0.05$; **$p < 0.01$; †$p < 0.001$

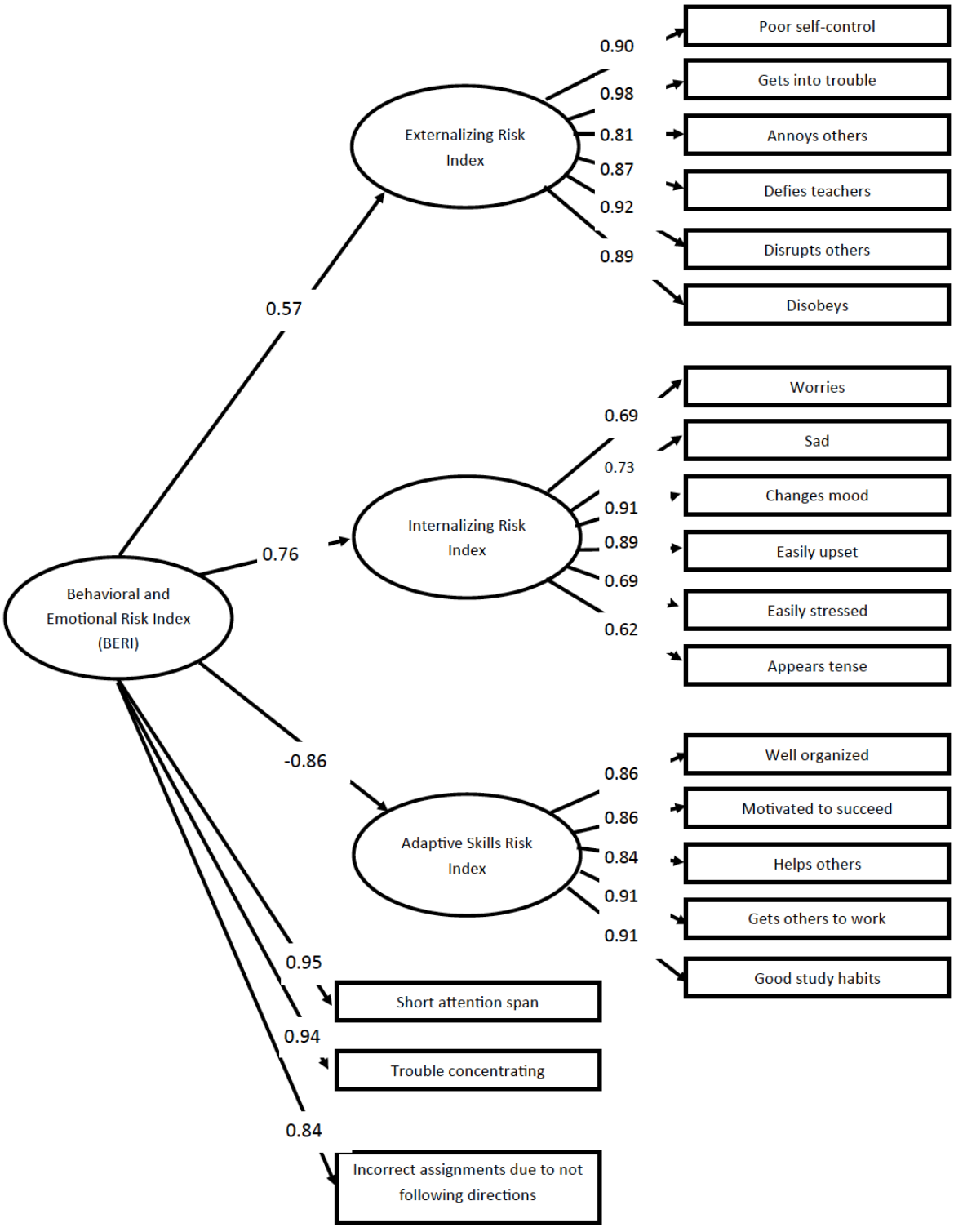

Figure 2. Standardized parameter estimates of the confirmatory factor analysis of the Somali kindergarten population.

Chapter 4

Synthesis and General Discussion

This multi-study dissertation first sought to systematically evaluate the current landscape of teacher-reported universal screening measures for the identification of students with SEB difficulties. The results of Study 1 demonstrated that the psychometric evidence reported in SEB rating scales' manuals is severely lacking both in terms of the quantity and quality of psychometric analyses. Specifically, the results of this study highlighted a clear deficit in the information reported pertaining to psychometrics in the rating scale manuals at time of publication, which was particularly evident in the area of bias evaluation. For this reason, Study 2 of this dissertation aimed to expand on the bias evaluation evidence of one of the reviewed universal screening measures, the Behavioral Assessment Scale for Children, Third Edition, Behavioral and Emotional Screening Scale (BASC-3 BESS) Teacher, by assessing its metric invariance—a critical first step in bias evaluation procedures—for a Somali kindergarten student population. While the results of Study 2 demonstrated adequate internal reliability and validity based on the internal structure for the study population, it also brought to light the extent to which rating scale manuals omit vital information related to both the analyses and results of psychometric evaluations. Therefore, the results of both studies emphasize the importance of thorough examination and reporting of psychometric evidence for all rating scales.

Implications for Research and Practice

Research. The results of this dissertation provide several research implications for review. First, while previous research provided thorough reviews of the most frequently used SEB rating scales for universal screening purposes (Brown, Scott-Little,

Amwake, & Wynn, 2007; Halle & Darling-Churchill, 2016; Jenkins et al., 2014; Niemeyer & Scott-Little, 2002; Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007; Yates et al., 2008), none have utilized a systematic approach to identify and review available universal screening rating scales specifically for the SEB component of kindergarten readiness. Further, limited research has explored the psychometric properties of rating scales with culturally diverse student populations. Thus, the primary research implication for this dissertation was to fill much-needed gaps in the literature by first identifying and thoroughly reviewing the psychometric properties of any available rating scale appropriate for use within universal screening procedures for the SEB component of kindergarten readiness, and second, conducting a bias evaluation study of the most frequently used SEB rating scale (Bruhn, Woods-Groves, & Huddle, 2014) specifically for culturally-diverse student population.

Overall, the results of this dissertation demonstrated a clear and consistent finding: limited or poor reporting of psychometric evidence for rating scales, both in terms of the quantity of psychometric properties reviewed for each rating scale at the time of manual publication, and the quality of reporting that is provided in the rating scale manual. Such omissions of reviewed psychometric properties make it difficult for the consumers to determine the applicability of the rating scales for their specific student populations and insure appropriate test interpretation and use. Further, such omissions in reporting negatively impact researchers who may attempt to replicate or compare findings of psychometric research to results of those reported for normative samples. This is particularly problematic as the technical manual of an assessment provides the foundation for appropriate interpretation and use. Test developers hold the responsibility

for ensuring that technical manuals contain clear descriptions of rating scales' purpose and psychometric evidence prior to disseminating the scales to general consumers (AERA et al., 2013).Therefore, it is imperative that test developers first make every effort to ensure that rating scales have undergone sufficient evaluation to fully support its use by following guidelines established by the Standards for Psychological and Educational Testing prior to its dissemination (AERA et al., 2013). Second, test developers must ensure appropriate specificity and clarity when reporting both analyses and results of psychometric testing. Such specificity and clarity may reduce misinterpretations of psychometric evidence and misuse of scales. Lastly, test developers must include explicit reports of information that is lacking or unavailable, while making efforts to periodically report newly available psychometric evidence over time. The results of this study then demonstrated the need of future research to provide thorough evaluations of the psychometric properties of rating scales specifically in the areas in which analyses or results are insufficiently reviewed in rating scale manuals.

It must be noted that this study demonstrated poor and inconsistent reporting of psychometric evidence in technical manuals of rating scales specifically because it opted to evaluate the state of the literature immediately following publication at the point in time in which they are readily available to consumers. As a result, the results of this dissertation provided evidence supporting the need for better reporting of psychometric evidence in rating scale manuals. Notably, additional psychometric evaluations may have been completed by independent researchers following the publication of the rating scale manuals that may provide further psychometric evidence for rating scales. Future researchers may then choose to provide comprehensive reviews of all psychometric

evaluations of rating scales by integrating psychometric reports of rating scales as provided by both rating scale manuals and independent researchers in the field. Future researchers may also choose to complete meta-analyses of all available psychometric results for each available rating scale, synthesizing results of all available studies, which may provide stronger evidence for the use of rating scales for specific student populations (Petticrew, & Roberts, 2006).

Lastly, the results of this study highlighted the significant gap in the literature specifically related to bias evaluations of rating scales. Most rating scales reviewed through this dissertation were severely lacking in bias evaluation reports. This is particularly problematic as schools in the United States are becoming increasingly more diverse (U.S. Department of Education, 2018). Further, exploring the psychometric properties of rating scales for diverse populations based solely on reported race may no longer be sufficient as research continues to demonstrate the impact of cultural factors on behavior (Braun et al., 2007; Hernandez, 2012). Therefore, future research must take into consideration the impact of both racial and cultural factors when conducting bias evaluation studies of rating scales.

Practice. The use of teacher rating scales for the identification of students who require additional SEB interventions has been demonstrated to be an effective and equitable practice in schools (Dever, Raines, Dowdy, & Hostutler, 2016; Raines, Dever, Kamphaus, & Roach, 2012). However, schools bear the responsibility of identifying the most appropriate rating scale for use with their specific student population. Unfortunately, schools oftentimes lack the resources – both in terms of time and cost – for thoroughly reviewing the available rating scales in order to identify the most

appropriate ones for use in their setting. This dissertation then serves as an important first step in aiding in the identification process of rating scales to be used as part of universal screening strategies. Schools may use the results of this dissertation to help identify the best rating scale for use based on their purpose for screening and student population.

The results of this dissertation, however, demonstrate the importance of being able to critically review and evaluate potential rating scales for use in their setting as quickly summarizing rating scale manuals may not be sufficient in the identification of rating scales. The results of this dissertation demonstrate that background knowledge in psychometric properties is essential in determining if rating scale manuals provide sufficient psychometric properties to support their use in specific school settings. School psychologists often have the requisite knowledge in measurement that would prove useful for such tasks. Schools are then urged to involve school psychologists in their search for the most appropriate rating scales to be used as part of universal screening procedures for students who may benefit from early SEB intervention services.

Conclusions

This dissertation provided a thorough review of rating scales that may be used as part of screening procedures for the identification of students who have difficulties with the SEB components of kindergarten readiness. In addition, this dissertation provided preliminary evidence for the use of the BASC-3 BESS Teacher, a frequently used SEB rating scale for screening purposes (Bruhn et al., 2014), with a Somali kindergarten population. However, results of this dissertation highlighted the lack of reporting of psychometric evidence for most rating scales, which was particularly prominent for bias evaluations. Therefore, future research must work to fill this gap by continuing to explore

the psychometric properties of available rating scales, with an emphasis on bias evaluation studies in order to demonstrate measurement invariance of rating scales with diverse student populations.

References

Abikoff, J., Courtney, M., Pelham, W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology, 21,* 519-533.

Abry, T., Latham, S., Bassok, D., & LoCasale-Crouch, J. (2015). Preschool and kindergarten teachers' beliefs about early school competencies: Misalignment matters for kindergarten adjustment. *Early Childhood Research Quarterly, 31,* 78-88. doi: http://dx.doi.org/10.1016/j.ecresq.2015.01.001

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

Albers, C. A., & Kettler, R. J. (2014). Best practices in universal screening. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology VI* (pp. 121-132). Bethesda, MD: National Association of School Psychology.

Alfonso, V. C., & Flanagan, D. P. (2009). Assessment of preschool children: A framework for evaluating the adequacy of the technical characteristics of norm-referenced instruments. In B. Mowder, F. Rubinson, & A. Yasik (Eds.), *Evidence based practice in infant and early childhood psychology* (pp. 129–166). New York, NY: Wiley.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association. Brown, Schott-Little, Amwake, & Wynn, 2007

Anastasi, A., & Urbina, S. (1997). *Psychological Testing Seventh Edition*. Minneapolis, MN: Pearson.

Barth, J. M., Dunlap, S. T., Dane, H., Lochman, J. E., & Wells, K. C. (2004). Classroom environment influences on aggression, peer relations, and academic focus. *Journal of School Psychology, 42*, 115-133. doi: 10.1016/j.jsp.2003.11.004

Bates, L. A., & Glick, J. E. (2013). Does it matter if teachers and schools match the student? Racial and ethnic disparities in problem behaviors. *Social Science Research, 42,* 1180-1190. doi: http://dx.doi.org/10.1016/j.ssresearch.2013.04.005

Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist, 57,* 111-127. doi: 10.1037//0003-066X.57.2.111

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology, 66,* 711-731. doi: 10.1146/annurev-psych-010814-015221

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78,* 647-663. Retrieved from http://www.jstor.org/stable/4139250

Bodrova, E., & Leong, D. J. (2008). Developing self-regulation in kindergarten: Can we keep all the crickets in the basket? *Young Children on the Web,* 1-3.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 4,* 313-326.

Bradshaw, C. P., Mitchell, M. M., O'Brennan, L. M., & Leaf, P. J. (2010). Multilevel exploration of factors contributing to the overrepresentation of Black students in office disciplinary referrals. *Journal of Educational Psychology, 102,* 508-520. doi: 10.1037/a0018450

Braun, L., Fausto-Sterling, A., Fullwiley, D., Hammonds, E. M., Nelson, A., Quivers, W., …, & Shields, A. E. (2007). Racial categories in medical practice: How useful are they? *PLOS Medicine, 4,* 1423-1428.

Breslau, J., Miller, E., Breslau, N., Bohnert, K., Lucia, V., & Schweitzer, J. (2009). The impact of early behavior disturbances on academic achievement in high school. *Pediatrics, 123,* 1472-1476. doi: 10.1542/peds.2008-1406

Brown, T. A. (2015). *Methodology in the social sciences. Confirmatory factor analysis for applied research (2nd ed.).* New York, NY: The Guilford Press.

Brown, G., Schott-Little, C., Amwake, L., & Wynn, L. (2007). *A review of methods and instruments used in state and local school readiness evaluations* (Issues & Answers Report, REL 2007–No. 004). Washington, DC: U.S. Department of Educa-tion, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educa-tional Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bruhn, A. L., Woods-Groves, S., & Huddle, S. (2014). A preliminary investigation of emotional and behavioral screening practices in K-12 schools. *Education and Treatment of Children, 37,* 611-634. doi: https://doi.org/10.1353/etc.2014.0039

Byrne, B. M. (2012). *Structural Equation Modeling with Mplus.* New York, NY: Routledge.

Campbell, J. M., & Hammond, R. K. (2014). Best practices in rating scale assessment of children's behavior. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology VI* (pp. 287-304). Bethesda, MD: National Association of School Psychology.

Capps, R., & Fix, M. (Eds.). (2012). *Young children of Black immigrants in America.* Washington, DC: Migration Policy Institute.

Carter, A. S. Briggs-Gowan, M. J., & Davis, N. O. (2004). Assessment of young children's social-emotional development and psychopathology: recent advances and recommendations for practice. *Journal of Child Psychology and Psychiatry, 45,* 109-134.

Choi, J. Y., Elicker, J., Christ, S. L., & Dobbs-Oates, J. (2016). Predicting growth trajectories in early academic learning: Evidence from growth curve modeling with Head Start children. *Early Childhood Research Quarterly, 36,* 244-258. doi: http://dx.doi.org/10.1016/j.ecresq.2015.12.017

Collaborative for Academic, Social, and Emotional Learning (2018). *Core SEL competencies.* Retrieved from https://casel.org/core-competencies/

Conners, C. K. (2009). *Conners Early Childhood.* North Tonawanda, NY: Multi-Health Systems.

Cook, C. R., Volpe, R. J., & Livanis, A. (2010). Constructing a roadmap for future universal screening research beyond academics. *Assessment for Effective Intervention, 35,* 197-205. doi: 10.1177/1534508410379842

Coolahan, K., Fantuzzo, J., Mendez, J., & McDermott, P. (2000). Preschool peer interactions and readiness to learn: Relationships between classroom peer play and learning behaviors and conduct. *Journal of Educational Psychology, 92,* 458-465. doi: I0.1037MXI22-0663.92.3.458

Cooper, D. H., & Farran, D. C. (1988). Behavioral risk factors in kindergarten. *Early Childhood Research Quarterly, 3,* 1-19.

Crosby, D. A., & Dunbar, A. S. (2012). Patterns and predictors of school readiness and early childhood success among young children in Black immigrant families. In R. Capps & M. Fix (Eds.). *Young children of Black immigrants in America* (pp. 183-228). Washington, DC: Migration Policy Institute.

Cullinan, D., & Epstein, M. H. (2013). *Emotional and Behavioral Screener (EBS).* Austin, TX: Proed.

Dawson, P. (2014). Best practices in assessing and improving executive skills. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology*. Bethesda, MD: National Association of School Psychology.

Darboe, K. (2003). New immigrants in Minnesota: The Somali immigration and assimilation. *Journal of Developing Societies, 19,* 458-472.

Datar**, A.,** & Gottfried, M. A. (2015). School entry age and children's social-behavioral skills: Evidence from a national longitudinal study of U.S. kindergartners. *Educational Evaluation and Policy Analysis, 37,* 333-353. doi: 10.3102/0162373714547268

Dever, B. V., Mays, K. L., Kamphaus, R. W., & Dowdy, E. (2012). The factor structure of the BASC-2 Behavioral and Emotional Screening System teacher form, child/adolescent. *Journal of Psychoeducational Assessment, 30,* 488-495. doi:10.1177/0734282912438869

Dever, B. V., Raines, T. C., Dowdy, E., & Hostutler, C. (2016). Addressing disproportionality in special education using a universal screening approach. *The Journal of Negro Education, 85,* 59-71.

Digital Promise Global (2016). The growing diversity in today's classroom. Retrieved from https://digitalpromise.org/wp-content/uploads/2016/09/lps-growing_diversity_FINAL-1.pdf

DiPerna, J. C., Bailey, C. G., & Anthony, C. (2014). Broadband screening of academic and social behavior. 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. 223-248). Washington, DC: American Psychological Association.

DiPerna, J. C., Volpe, R. J., & Eliott, S. N. (2002). A model of academic enablers and mathematics achievement in the elementary grades. *Journal of School Psychology, 43,* 379-392. doi: 10.1016/j.jsp.2005.09.002

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43,* 121-149. doi: 10.1177/0748175610373459

Division of Early Childhood Development (2016). *The 2015-2016 kindergarten readiness assessment report*. Baltimore, MD: Division of Early Childhood Development

Dowdy, E., Kamphaus, R. W., Twyford, J. M., & Dever, B. V. (2014). Culturally competent behavioral and emotional screening. In M. D. Weist, N. A. Lever, C. P. Bradshaw, & J. S. Owens (Eds.). *Handbook of school mental health: Research, training, practice, and policy* (pp. 311-321). New York, NY: Springer Science+Business Media.

Dowdy, E., Twyford, J. M., Chin, J. K., DiStefano, C. A., Kamphaus, R. W., & Mays, K. L. (2011). Factor structure of the BASC-2 Behavioral and Emotional Screening System student form. *Psychological Assessment, 23,* 379-387. doi:10.1037/a0021843

Downer, J. T., Goble, P., Myers, S. S., & Pianta, R. C. (2016). Teacher-child racial/ethnic match within pre-kindergarten classrooms and children's early school adjustment. *Early Childhood Research Quarterly, 37,* 26-38. doi: http://dx.doi.org/10.1016/j.ecresq.2016.02.007

Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education, 77,* 267-282.

Edossa, A. K., Schroeders, U., Weinert, S., & Artelt, C. (2018). The development of emotional and behavioral self-regulation and their effects on academic achievement in childhood. *International Journal of Behavioral Development, 42,* 192-202. doi: 10.1177/0165025416687412

Eklund, K., Renshaw, T. L., Dowdy, E., Jimerson, S. R., Hart, S. R., Hones, C. N., & Earhart, J. (2009). Early identification of behavioral and emotional problems in youth: Universal screening versus teacher-referral identification. *The California School Psychologist, 14,* 89-95.

Epstein, M. (1998). *Behavioral and Emotional Rating Scale 2 (BERS-2).* Austin, TX: Proed.

Epstein, M. (1998). Using Strength-Based Assessment in Programs for Children with Emotional and Behavioral Disorders. *Beyond Behavior, 9*(2), 25-27. Retrieved from http://www.jstor.org.ezp1.lib.umn.edu/stable/44709177

Epstein M.H. et al. (2003) Strength-Based Approaches to Assessment in Schools. In: Weist M.D., Evans S.W., Lever N.A. (eds) *Handbook of school mental health advancing practice and research: Issues in clinical child psychology* (pp. 285-299). Springer, Boston, MA

Essex, M. J., Kraemer, H. C., Slattery, M. J., Burk, L. R., Boyce, W. T., Woodward, H. R., & Kupfer, D. (2009). Screening for childhood mental health problems: Outcomes and early identification. *Journal of Child Psychology and Psychiatry, 50,* 562-570. doi: 10.1111/j.1469-7610.2008.02015.x

Fan, X. & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational Psychology Review, 13*(1), 1-22. doi: 1040-726X/01/0300-0001

Ferguson, R. F. (1998). Teachers' perceptions and expectations and the Black-White test score gap. *Urban Education, 38,* 460-507. doi: 10.1177/0042085903254970

Fish, R. E. (2017). The racialized construction of exceptionality: Experimental evidence of race/ethnicity effects on teachers' interventions. *Social Science Research, 62,* 317-334. doi: http://dx.doi.org/10.1016/j.ssresearch.2016.08.007

Floyd, R. G., Shands, E. I., Alfonson, V. C., Phillips, J. F., Autry, B. K., Mosteller, J. A., …, Irby, S. (2015). A systematic review and psychometric evaluation of adaptive behavior scales and recommendations for practice. *Journal of Applied School Psychology, 31,* 83-113. doi: 10.1080/15377903.2014.979384

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2015). *Behavior Rating Inventory of Executive Function, Second Edition (BRIEF2).* Lutz, FL: PAR.

Gilliam, W. S., Maupin, A. N., Reyes, C. R., Accavitti, M., & Shic, F. (2016). Research Study Brief on the Yale Child Study Center. Do early educators' implicit biases regarding sex and race relate to behavior expectations and recommendations of preschool expulsions and suspensions? Retrieved from https://medicine.yale.edu/childstudy/zigler/publications/Preschool%20Implicit%20Bias%20Policy%20Brief_final_9_26_276766_5379_v1.pdf

Glover, 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., & Elliott, S. N. (2008). *Social Skills Improvement System (SSiS).* Minneapolis, MN: Pearson.

https://www.jstor.org/stable/20152704

Halle, T. G., & Darling-Churchill, K. E. (2016). Review of measures of social and emotional development. *Journal of Applied Development Psychology, 45,* 8-18. doi: http://dx.doi.org/10.1016/j.appdev.2016.02.003

Halperin, J. M., & McKay, K. E. (2008). *Children's Aggression Scale (CAS).* Lutz, FL: PAR.

Hernandez, D. J. (2012). Young children in Black immigrant families from Africa and the Caribbean. In R. Capps & M. Fix (Eds.). *Young children of Black immigrants in America* (pp. 75-118). Washington, DC: Migration Policy Institute.

Hightower, A. D., & Perkins, P. E. (2010). *Teacher-Child Rating Scale 2.1 (T-CRS 2.1).* Rochester, NY: Children's Institute, Inc.

Hogan, T. P. (2017). [Test review of BASC-3 Behavioral and Emotional Screening System]. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.). *The twentieth mental measurements yearbook* (pp. 61-64). Lincoln, NE: Buros Center for Testing.

Hosp, J. L., & Reschly, D. J. (2003). Referral rates for intervention or assessment: A meta-analysis of racial differences. *The Journal of Special Education, 37,* 67-80.

Howse , R. B., Calkins, S. B. , Anastopoulos , A. D., Keane, S. P. & Shelton, T. L. (2003). Regulatory contributors to children's kindergarten achievement, Early *Education and Development, 14,* 101-120, doi: 10.1207/s15566935eed1401_7

Howse, R. B., Lange, G., Farran, D. C., & Boyles, C. D. (2010). Motivation and self-regulation as predictors of achievement in economically disadvantaged young children. *The Journal of Experimental Education, 71,* 151-175. Retrieved from

Hu, L. & Bentler, P. M. (1990). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1-55. doi:10.1080/107055199095

IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Ikeda, M. J., Neessen, E., & Witt, J. C. (2008). Best practices in universal screening. In A. Thomas & J. Grimes (Eds), *Best practices in school psychology V* (pp. 721-734). Bethesda, MD: National Association of School Psychologists.

Jenkins, L. N., Demaray, M. K., Wren, N. S., Secord, S. M., Lyell, K. M., Magers, A. M., …, Tennant, J. (2014). A critical review of five commonly used social-emotional and behavioral screeners for elementary or secondary schools. *Contemporary School Psychology*. doi: 10.1007/s40688-014-0026-6

Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health, 105,* 2283-2290.

Kagan, S. L., Moore, E., & Bradekamp, S. (1995). *Reconsidering children's early development and learning: Towards common views and vocabulary.* Washington, DC: U.S. Government Printing Office.

Kamphaus, R. W., & Reynolds, C. R. (2015). *BASC3 Behavioral and Emotional Screening System (BESS).* Minneapolis, MN: Pearson.

Kamphaus, R. W., Reynolds, C. R., & Dever, B. V. (2014). Behavioral and mental health screening. 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. 249-273). Washington, DC: American Psychological Association.

Kelley, K. & Pornprasertmanit, S. (2016). Confident intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21,* 69-92.

King, K., 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 Pschoeducational Assessment, 30,* 527-238. doi: 10.1177/0734282912440673

Kiperman, S., Black, M. S., McGill, T. M., Harrell-Williams, L. M., & Kamphaus, R. W. (2014). Predicting Behavior Assessment System for Children—Second edition self-report of personality child form results using the Behavioral and Emotional Screening System student form: A replication study with an urban predominantly Latino/a sample. *Journal of Pscyhoeducational Assessment, 32,* 587-596. doi:10.1177/0734282914529200

Koch, J. M. (2007). How schools can best support Somali students and their families. *International Journal of Multicultural Education, 9*(1), 1-15.

Lewit, E. M., & Baker, L. S. (1995). School readiness. *The Future of Children, 5,* 128-139. Retrieved from http://www.jstor.org/stable/1602361

Li-Grining, C. P., Vortruba-Drzal, E., Maldonado-Carreno, C., & Haas, K. (2010). Children's early approaches to learning and academic trajectories through fifth grade. *Developmental Psychology, 46,* 1062-1077. doi: 10.1037/a0020066

Liew, J. (2012). Effortful control, executive functions, and education: Bringing self-regulatory and social-emotional competencies to the table. *Child Development Perspectives, 6,* 105-111. doi: 10.1111/j.1750-8606.2011.00196.x

Linder, S. M., Ramey, M. D., & Zambak, S. (2013). Predictors of school readiness in literacy and mathematics: A selective review of the literature. *Early Childhood Research & Practice, 15*(1). Retrieved from http://ecrp.uiuc.edu/v15n1/linder.html

Lin, H., Lawrence, F. R., & Gorrell, J. (2003). Kindergarten teachers' views of children's readiness for school. *Early Childhood Research Quarterly, 18,* 225-237. doi: https://doi.org/10.1016/S0885-2006(03)00028-0

Lopes, P. N., Salovey, P., Cote, S., & Beers, M. (2005). Emotion regulation abilities and the quality of social interaction. *Emotion, 5,* 113-118. doi: 10.1037/1528-3542.5.1.113

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130-149.

Major, S. O., & Seabra-Santos, M. J. (2015). Are we talking about the same child? Parent-teacher ratings of preschoolers' social-emotional behaviors. *Psychology in the Schools, 52,* 789-799. doi: 10.1002/pits.21855

McClellan, M. M., & Cameron, C. E. (2011). Self-regulation and academic achievement in elementary school children. In R. M. Lerner, J. V. Lerner, E. P. Bowers, S. Lewin-Bizan, S. Gestsdottir, & J. B. Urban (Eds.), *Thriving in childhood and adolescence: The role of self-regulation processes. New directions for child and adolescent development, 133,* 29-44.

Merrell, K. W. (2002). *School Social Behavior Scales Second Edition (SSBS-2).* Baltimore, MD: Paul H Brookes Publishing Co.

Merrell, K. W. (2010). Commentary: Better methods, better solutions: Developments in school-based behavioral assessment. *School Psychology Review, 39,* 422-426.

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: http://dx.doi.org/10.1037/spq0000085

Mollborn, S. (2016). Young children's developmental ecologies and kindergarten readiness. *Demography, 53,* 1853-1882. doi: 10.1007/s13524-016-0528-0

Monaghan, W. (2006). The facts about subscores. *R & D Connections*. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

Montes, G., Lotcyzewski, B. S., Halterman, J. S., & Hightower, A. D. (2012). School readiness among children with behavior problems at entrance into kindergarten: results from a US national study. *European Journal of Pediatrics, 171,* 541-548. doi: 10.1007/s00431-011-1605-4

Muthén, L.K., & Muthén, B.O. (2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén

Nagle, R. J. (2000). Issues in preschool assessment. In B. A. Bracken (Ed). *The psychoeducational assessment of preschool children.* Mahwah, NJ: Routledge.

Naglieri, J. A., LeBuffe, P. A., & Shapiro, V. B. (2011). *Devereux Student Strengths Assessment Mini (DESSA-Mini).* Lewisville, NC: Kaplan.

National Center for Education Statistics (2013). Table 136. Number and percentage distribution of kindergarteners, by fall 2010 kindergarten entry status and selected child, family, and school characteristics: 2010-111. In U.S. Department of

Education, National Center for Education Statistics (Ed.). *Digest of Education Statistics* (2013 ed.). Retrieved from https://nces.ed.gov/programs/digest/d12/tables/dt12_136.asp

National Center for Education Statistics (2018). Table 203.60: Enrollment and percentage distribution of enrollment in public elementary and secondary schools, by race/ethnicity and level of education: Fall 1998 through fall 2023. In U.S. Department of Education, National Center for Education Statistics (Ed.), *Digest of Education Statistics* (2018 ed.). Retrieved from https://nces.ed.gov/programs/digest/d17/tables/dt17_203.50.asp?current=yeshttps://nces.ed.gov/programs/digest/d17/tables/dt17_203.50.asp?current=yes

National Education Goals Panel (1997). *Getting a good start in school.* Washington, DC: National Education Goals Panel. Retrieved from https://govinfo.library.unt.edu/negp/reports/good-sta.htm

Niemeyer, J., & Scott-Little C. (2002). *Assessing kindergarten children: A compendium of assessment instruments.* Greensboro: University of North Carolina.

Oakes, W. P., Lane, K. L., Cox, M. L., & Messenger, M. (2014). Logistics of behavior screenings: How and why do we conduct behavior screenings at our school? *Preventing School Failure: Alternative Education for Children and Youth, 58,* 159-170. doi: 10.1080/1045988X.2014.895572

Oakes, W. P., Lane, K. L., & Ennis, R. P. (2016). Systematic screening at the elementary level: Considerations for exploring and installing universal behavior screening. *Journal of Applied School Psychology, 32,* 214-233. doi: 10.1080/15377903.2016.1165325

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

Oakes, W. P., Lane, K. L., Cox, M. L., & Messenger, M. (2014). Logistics of behavior screenings: How and why do we conduct behavior screenings at our school? *Preventing School Failure: Alternative Education for Children and Youth, 58,* 159-170. doi: 10.1080/1045988X.2014.895572

Oakes, W. P., Lane, K. L., & Ennis, R. P. (2016). Systematic screening at the elementary level: Considerations for exploring and installing universal behavior screening. *Journal of Applied School Psychology, 32,* 214-233. doi: 10.1080/15377903.2016.1165325

Parisi, D. M., Ihlo, T., & Glover, T. A. (2014). Screening within a multitiered early prevention model: Using assessment to inform instruction and promote students' response to intervention. In R J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings* (pp. 19-46). Washington, DC: American Psychological Association.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide.* Malden, MA: Blackwell Publishing

Pigott, R. L., & Cowen, E. L. (2000). Teacher race, child race, racial congruence, and teacher ratings of children's school adjustment. *Journal of School Psychology, 38,* 177-196.

Quirk, M., Grimm, R., Furlong, M. J., Nylund-Gibson, K., & Swami, S. (2016). The association of Latino children's kindergarten school readiness profiles with grade 2-5 literacy achievement trajectories. *Journal of Educational Psychology, 108,* 814-829. doi: http://dx.doi.org.ezp1.lib.umn.edu/10.1037/edu0000087

Raver, C. C., Garner, P. W., & Smith-Donald, R. (2007). The roles of emotion regulation and emotion knowledge for children's academic readiness: Are the links causal? In R. C. Pianta, M. J. Cox, & K. L. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 121- 147). Baltimore, MD, US: Paul H Brookes Publishing.

Raaijmakers, M. A. J., Smidts, D. P., Sergeant, J. A., Maassen, G. H., Posthumus, J. A., van Engeland, H., & Matthys, W. (2008). Executive functions in preschool children with aggressive behavior: Impairments in inhibitory control. *Journal of Abnormal Child Psychology, 36,* 1097-1107. doi: 10.1007/s10802-008-9235-7

Raines, T. C., Dever, B. V., Kamphaus, R. W., & Roach, A. T. (2012). Universal screening for behavioral and emotional risk: A promising method for reducing disproportionate placement in special education. *The Journal of Negro Education, 81,* 283-296.

Rastogi, S., Johnson, T. D., Hoeffel, E. M., & Dewery, M. P. (2011, September). *Census Briefs: The Black population: 2010.*

Ray, K., & Smith, M C. (2010). The kindergarten child: What teachers and administrators need to know to promote academic success in all children. *Early Childhood Education, 38,* 5-18. doi: 10.1007/s10643-010-0383-3

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

Reynolds, C.R., Kamphaus, R.W. (2015). *Behavior assessment for children: Third edition*. (BASC-3). Bloomington, MN: Pearson.

Reynolds, C. R., & Carson, A. D. (2005). Methods for assessing cultural bias in tests. In C. Frisby & C. R. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 795–823). Hoboken, NJ: Wiley.

Rimm-Kaufman, S. E., Pianta, R. C., & Cox, M. J. (2000). Teachers' judgements of problems in the transition to kindergarten. *Early Childhood Research Quarterly, 15,* 147-166.

Rios J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26,* 108-116. doi: 0.7334/psicothema2013.260

Roll, J., Koglin, U., & Petermann, F. (2012). Emotion regulation and childhood aggression: Longitudinal associations. *Child Psychiatry and Human Development, 43,* 909-923. doi: 10.1007/s10578-012-0303-4

Raines, T. C., Dever, B. V., Kamphaus, R. W., & Roach, A. T. (2012). Universal screening for behavioral and emotional risk: A promising method for reducing disproportionate placement in special education. *The Journal of Negro Education, 81,* 283-296.

Saas, D. A., & Schmitt, T. A. (2013). Testing measurement and structural invariance. In T. Teo (Ed.). *Handbook of quantitative methods for educational research* (pp.315-345). Boston, MA: Sense Publishers.

Sabol, T. J., & Pianta, R. C. (2012). Patterns of school readiness forecast achievement and socioemotional development at the end of elementary school. *Child Development, 83,* 282-299. doi: 10.1111/j.1467-8624.2011.01678.x

Scuglik, D. L., Alarcón, R. D., Lapeyre, A. C., Williams, M. D., & Logan, K. M. (2007). When the poetry no longer rhymes: Mental health issues among Somali immigrants in the USA. *Transcultural Psychiatry, 44,* 581-595. doi:10.1177/1363461507083899

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

Salvia, J., & Ysseldyke, J. (1998). *Assessment Seventh Edition*. Boston, MA: Houghton Mifflin Company.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107-120.

Sink, C. A., & Carlisle, K. L. (2017). [Test review of BASC-3 Behavioral and Emotional Screening System]. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.). *The twentieth mental measurements yearbook* (pp. 64-67). Lincoln, NE: Buros Center for Testing.

Skiba, R. J., Poloni-Staudinger, L., Gallini, S., Simmons, A. B., & Feggins-Azziz, R. (2006). Disparate access: The disproportionality of African American students with disabilities across educational environments. *Council for Exceptional Children, 72,* 411-424.

Smith-Millman, M. K., Flaspohler, P. D., Maras, M. A., Splett, J. W., Kristy Warmbold, Dinnen, H. & Luebbe, A. (2017). Differences between teacher reports on universal risk assessments. *Advances in School Mental Health Promotion*, *10,* 235-249. doi: 10.1080/1754730X.2017.1333914

Snow, K. L. (2006). Measuring school readiness: Conceptual and practical considerations. *Early Education and Development, 17,* 7-41.

Sullivan, A. L., Houri, A. K., & Sadeh, S. (2016). Demography and early academic skills of students from immigrant families: The kindergarten class of 2011. School Psychology Quarterly, 31, 149-162. doi: 10.1037/spq0000137

Tenenbaum, H. R., 7 Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology, 99,* 253-273. doi:10.1037/0022-0663.99.2.253

Trentacosta C. J., & Izard, C. E. (2007). Kindergarten children's emotion competence as a predictor of their academic competence in first grade. *Emotion, 7,* 77-88. doi: 10.1037/1528-3542.7.1.77

U.S. Department of Education (2018). The condition of education, 2018. Retrieved from https://nces.ed.gov/pubs2018/2018144.pdf

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología, 33,* 755-782.

Volpe, R. J., Briesch, A. M., & Chafouleas, S. M. (2010). Linking screening for emotional and behavioral problems to problem-solving efforts: An adaptive

model of behavioral assessment. *Assessment for Effective Intervention, 35*, 240-244. doi*: 10.1177/1534508410377194*

Walker, H. M., Severson, H. H., & Feil, E. G. (2014). *Systematic Screening for Behavior Disorders (SSBD 2nd Ed.).* Eugene: OR. Pacific Northwest Publishing.

Welsh, R. O., & Little, S. (2018). The school discipline dilemma: A comprehensive review of disparities and alternative approaches. *Review of Educational Research, 88,* 752-794. doi: 10.3102/0034654318791582

West, J., Hausken, E. G., & Collins, M. (1993). Readiness for kindergarten: Parent and teacher beliefs. *National Center for Education Statistics.* Retrieved from https://files.eric.ed.gov/fulltext/ED363429.pdf

Wiesner, M., & Schanding, G. T. (2013). Exploratory structural equation modeling, bifactor models, and standard confirmatory factor analysis models: Application to the BASC-2 Behavioral and Emotional Screening System Teacher Form. *Journal of School Psychology, 51,* 751-763.

Yates, T., Ostrosky, M. M., Cheatham, G. A., Fettig, A., Shaffer, L., & Santos, R. M. (2008). Research synthesis on screening and assessing social-emotional competence. *The Center on the Social and Emotional Foundations for Early Learning.* Retrieved from http://csefel.vanderbilt.edu/documents/rs_screening_assessment.pdf

Zelazo, P. D., Qu, L., & Kesek, A. C. (2010). Hot executive function: Emotion and the development of cognitive control. In S. D. Calkins & M. A. Bell (Eds.), *Child development at the intersection of emotion and cognition* (pp. 97-111). Washington, DC: American Psychological Association.

Zelazo, P. D., & Müller, U. (2002). Executive function in typical and atypical development. In U. Goswami (Ed), *Blackwell handbook of childhood cognitive development* (pp. 445-469). Hoboken, NJ: Blackwell Publishers Ltd.