Technical Adequacy of the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale: A Systematic Review, Quantitative Synthesis, & Measurement Invariance Study

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Annie K. Goerdt

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Advisor: Faith G. Miller, Ph.D.

May 2023

Land Acknowledgement

We acknowledge that the University of Minnesota Twin Cities is built within the traditional homelands of the Dakota people. It is important to acknowledge the peoples on whose land we live, learn, and work as we seek to improve and strengthen our relations with our tribal nations. We also acknowledge that words are not enough. We must ensure that our institution provides support, resources, and programs that increase access to all aspects of higher education for our American Indian students, staff, faculty, and community members.

Positionality Statement

In research and scholarship, it is critical to acknowledge one's positionality. As defined by Arora et al. (2022), positionality refers to "All aspects of scholarly endeavors … influenced by the assumptions, attitudes, experiences, values, and sociopolitical location of the researcher" (p. 2). As the author of this dissertation, I represent a number of dominant identities (e.g., White, middle class, able-bodied), which inevitably impact my viewpoints and perspectives. Specifically, the following studies represent a quantitative approach to research with implications drawn from a singular White perspective. Racism is complex and deep-seated; it cannot be quantified. A number of critical theory principles (i.e., QuantCrit) were applied to this dissertation, including the critical questioning of prior data collection and analysis, transparency in reporting, and "going beyond the existing and over-used metrics" (Castillo & Gillborn, 2022). However, limitations are present throughout, including the categorization of racial or ethnic identity. These limitations are presented with transparency and discussed throughout the manuscript. In an effort for self-reflexivity, I aim to engage a more comprehensive QuantCrit lens to my future engagement in quantitative research methods, with the hope of centering the voices of marginalized groups.

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Faith Miller. Your ongoing support and guidance throughout my graduate career is immeasurable. The time and commitment you dedicated to my research milestones and experiences have been integral to my professional development. To my mentor, Dr. Elyse Farnsworth, thank you for providing a safe and welcoming space to myself and all graduate students. Your encouragement was instrumental to my growth as a practitioner and researcher. To my committee, Dr. Danielle Dupuis, Dr. Glenn Roisman, and Dr. Annie Hansen-Burke, thank you for providing your collective knowledge and guidance. I would not have completed this dissertation without your feedback and participation. To my parents and sister, your ongoing and unconditional support means the world to me. Words cannot express my gratitude for all you have done throughout my graduate career. Last but certainly not least, Dan, I would not be here without you. Thank you for pushing me, believing in me, and supporting me. *We* did it!

Abstract

Universal screening in educational settings provides an opportunity to identify students who may benefit from targeted support. In recent years, there has been an increased awareness of the need for universal screening in the social, emotional, and behavioral domains. Yet, the research base for these screening tools is scarce given its relative novelty, particularly in comparison to universal screening tools within academic domains. Therefore, this dissertation was guided and conceptualized by a need to begin filling this gap in the literature. Specifically, these studies examined the technical adequacy of one widely used measure, the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS). Broadly, the purpose of this multi-study dissertation was two-fold: (a) systematically review and quantitatively synthesize the existing evidence of validity related to the SAEBRS-TRS; and (b) examine the extent of potential measurement invariance of the tool across the student characteristics of racial and ethnic identity, sex assigned at birth, and eligibility for free or reduced-price lunch. Results of the systematic review provide preliminary evidence of the SAEBRS-TRS technical adequacy across sources of validity evidence. Yet, more research is needed to gather a stronger research base for using and interpreting the SAEBRS-TRS scores across a variety of student characteristics. Results of the measurement invariance study provide initial support for the use of the SAEBRS-TRS across broad student characteristics of racial and ethnic identity, sex assigned at birth, and eligibility for free or reduced-price lunch. Inherent limitations of these studies, such as the categorization and homogenization of student identities, are presented. Future research must continue to move this work forward, further examining the SAEBRS-TRS and similar universal screening tools for the social, emotional, and behavioral domains.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The social, emotional, and behavioral (SEB) development of children and adolescents is critical for the overall well-being and functioning of youth. In our current context, the COVID-19 pandemic combined with continued acts of racial injustice, racial violence, and social unrest across the nation will undeniably result in increased needs across the SEB domains. Although there are a number of ways children and adolescents can be identified for SEB support, many continue to fall through the cracks (Merikangas et al., 2010). In the absence of systematic prevention, early identification, and evidence-based intervention, youth exhibiting SEB problems or at-risk for SEB problems are less likely to graduate and more likely to experience suspension, expulsion, truancy, and academic underachievement (Bradley et al., 2008). Fortunately, there is evidence that early identification combined with evidence-based intervention can ameliorate the likelihood of these negative outcomes.

As a gateway to the provision of SEB support, educational settings have a marked opportunity to engage in the adoption and implementation of promising methods to identify students in need. One method of early identification, universal screening, is performed in schools to identify students at-risk for SEB needs. Universal screening is defined as the process of systematically evaluating all students for academic and SEB difficulties who may be in need of additional services (Glover & Albers, 2007). Methods of conducting universal screening are varied, and can consist of structured referrals/nominations, observations, and rating scales.

Despite the promise of universal screening to serve as a critical conduit to responsive service delivery, these assessments may not perform equitably across students from different backgrounds or lived experiences, leading to inappropriate or unsound decisions. As we confront a national reckoning on issues of racial injustice and inequity, it is more important than ever that we critically evaluate potentially inequitable practices. To this point, the proposed studies are particularly salient in evaluating the usability and interpretability of a widely used universal screening tool across student characteristics. Specifically, the primary aim of this multi-study dissertation is to systematically review extant validity evidence and examine the measurement invariance of a popular universal screener, the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS; Kilgus et al., 2016).

Background

School-based universal screening in the SEB domains offers a promising option to preemptively identify students in need of SEB support, in hopes to ameliorate the risk of future negative outcomes. As such, researchers have developed numerous measures for the purpose of accurate and efficient universal SEB screening. Several frequently used measures include the Behavioral and Emotional Screening System, Third Edition (BESS-3; Kamphaus & Reynolds, 2015), the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), the Systematic Screening for Behavior Disorders (SSBD; Severson & Walker, 1992), and the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS; Kilgus et al., 2016). Although these measures have many strengths, they also have weaknesses that may limit their utility. For example, a systematic review of universal screeners indicated the SDQ displays limited validity evidence and the BESS-3 and SSBD are lengthy and time-consuming (Jenkins et al., 2014). Furthermore, bias evaluation studies (e.g., measurement invariance) for the teacher-report versions of these measures are relatively lacking.

Although universal screening represents a substantial improvement over other common identification methods, the defensibility of this approach rests mainly on the psychometric qualities and appropriate use of the measures. An established method of critically reviewing and evaluating the validity of these tools can occur by applying the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014). Specifically, validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association et al., 2014, p. 11). According to the *Standards*, there are numerous sources of validity evidence that encompass multiple aspects of a measure. The main sources of validity evidence include test content, response processes, internal structure, relations to other variables, and consequences of testing.

As part of this process, assessments must be evaluated for measurement bias. Although universal screening tools may permit us to make accurate inferences about the skills students possess, items on these assessments or the assessments themselves may be biased, leading to inappropriate or unsound decisions. This phenomenon, referred to as measurement bias, is defined as "components of test scores that differentially affect the performance of test takers and consequently the reliability/precision and validity of interpretations and uses of their test scores" (American Educational Research Association et al., 2014, p. 216). Therefore, studies of measurement bias are critical for examining the interpretation and use of an assessment with diverse student populations.

Some universal screening tools have been evaluated for measurement invariance, such as the BESS-3 (Kamphaus & Reynolds, 2015). However, the majority of studies examined the previous version of the measure (i.e., BESS-2; Dowdy et al., 2011; Splett et al., 2017). Nonetheless, one study examined the revised BESS-3, demonstrating invariance across the student-level variables of racial or ethnic identity, grade, and sex assigned at birth (Splett et al., 2020). This study also has limitations, as it was conducted within a single school district in the context of four elementary schools. Although these findings provide preliminary support, robust evidence to support measurement invariance commensurate with the widespread use of the measure is clearly lacking.

Another popular SEB screening tool, the Student Risk Screening Scale (SRSS; Drummond, 1994), has been minimally examined for measurement invariance. A single study of the SRSS indicated invariance across time and sex assigned at birth; yet important characteristics such as racial or ethnic identity were not reported and the study was conducted in a single, rural school district (Fredrick et al., 2019). Considering the importance of comparable functioning of scores across groups, this lack of prioritization in the examination of measurement invariance is both concerning and problematic.

The Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS)

The SAEBRS-TRS is included with the FastBridge suite of assessments and is a brief 19-item rating scale. The purpose of the SAEBRS-TRS is to identify students at-risk for social, academic, and emotional behavior problems. The SAEBRS-TRS consists of one full scale score (Total Behavior [TB]) comprised of three subscales (Social Behavior [SB], Academic Behavior [AB], Emotional Behavior [EB]). The developers originally established risk ranges in comparison to similar behavior rating scales, including the Social Skills Improvement System (Gresham & Elliot, 2008) and the Behavioral and Emotional Screening System, Second Edition (BESS-2; Kamphaus & Reynolds, 2007).

In interpreting SAEBRS-TRS scores, the developers initially recommended using cut scores to determine risk, with scores below a certain threshold meeting criteria for *At-Risk* or *Not*

At-Risk. Specifically, scores falling in the *At-Risk* range were 36 or below for Total Behavior, 12 or below for Social Behavior, 9 or below for Academic Behavior, and 17 or below for Emotional Behavior. However, in summer of 2021, a national norm-referenced model was released. The normative sample included 687 K-12 schools across 29 states, with the authors reporting the sample represented a similar composition to the national student population. Authors noted three specific student characteristics used for matching: racial or ethnic identity, sex assigned at birth, and eligibility for free or reduced-price lunch. Separate analyses (e.g., mixed-effects linear models) were conducted to ensure effects of racial or ethnic identity and free or reduced-price lunch rates did not have a significant effect on SAEBRS-TRS scores.

The norm-referenced model resulted in the following norms: *Low Risk* scores above the 16th national percentile, *Some Risk* scores between the third and 16th national percentile, and *High Risk* scores below the third national percentile. Revised cut scores (i.e., benchmarks) were developed based on national norms representing approximately one and two standard deviations below the mean. Scores falling in the *Some Risk* range were 24 to 36 or below for Total Behavior, 8 to 12 for Social Behavior, 6 to 9 for Academic Behavior, and 12 to 15 for Emotional Behavior. Scores falling in the *High Risk* range were 23 or below for Total Behavior, 7 or below for Social Behavior, 5 or below for Academic Behavior, and 11 or below for Emotional Behavior. These revisions are informative; however, results from studies conducted prior to summer of 2021 should be interpreted with caution as they were analyzed and interpreted using the previous cut scores. Importantly, in the SAEBRS-TRS technical report explaining these norms and revised benchmarks, the authors provided the following statement:

> It is important to note that a student’s score on the SAEBRS and mySAEBRS should never be used as the sole determinant of overall risk or intervention services. Instead,

these scores should be examined by a team consisting of the student's teacher(s), counselor, psychologist, administrative leader, and others who know the student well. SAEBRS and mySAEBRS must be compared with other sources of information about the student's behaviors to confirm the presence of risk and need for support (Illuminate Education, 2021).

Rationale

Universal screening is an integral component of a multi-tiered system of support (MTSS), which is a comprehensive framework that prioritizes alignment and allocation of support for student academic, social, emotional, and behavioral success. Universal screening allows educators to engage in data-based decision-making and early identification of students who may benefit from SEB support (Glover & Albers, 2007). Research has demonstrated that universal screening completed via rating scales provides a more comprehensive approach than other commonly used methods, such as teacher referral or Office Discipline Referrals (Miller et al., 2015). The SAEBRS-TRS is one universal screening tool used to systematically assess across the SEB domains. Yet, no quantitative synthesis of the SAEBRS-TRS literature exists. Additionally, only one study has examined the measurement invariance of the SAEBRS-TRS. However, the study combined the student- and teacher-report versions (i.e., mySAEBRS and SAEBRS-TRS), thus obscuring the scrutiny of the teacher-report version (von der Embse et al., 2019).

Moreover, the majority of research has been conducted by the developers of the measure (i.e., Kilgus, von der Embse, and colleagues). Research must be conducted outside of the group of developers to examine the psychometric defensibility of the SAEBRS-TRS in a variety of settings. Given the widespread use of the SAEBRS-TRS within our geographic locale, we became interested in understanding psychometric evidence to support the use of the measure, including the extent of measurement invariance. Therefore, the broad purpose of this multi-study dissertation was two-fold: (a) to synthesize empirical research related to the technical adequacy of the SAEBRS-TRS; and (b) to examine the measurement invariance of the SAEBRS-TRS across student characteristics.

Study 1: SAEBRS-TRS Systematic Review & Meta-Analysis

Systematic reviews and meta-analyses of validity evidence are vital considering the contemporary framework of validity evidence, in which the strength of a validity argument is cumulative and contingent on the extent of validity evidence (American Educational Research Association et al., 2014). In accordance with this modern understanding of validity and given the growing use of the SAEBRS-TRS in educational settings, this study will address a gap in the literature by presenting a quantitative, holistic synthesis of available validity evidence. Additionally, this study will provide context regarding the populations studied, such as grade level and racial or ethnic identity, and highlight the conditions under which the measure was administered. This is a critical component of a measure's psychometric defensibility, as validity evidence is dependent on the populations and settings in which measures are tested to determine applicability. Overall, the aim of Study 1 is to systematically review and quantitatively synthesize the available empirical research on the psychometric properties of the SAEBRS-TRS. As such, the results of Study 1 seek to address the following research questions:

1. What evidence of validity currently exists for the SAEBRS-TRS (e.g., response processes, test content, internal structure, relations to other variables, consequences of testing)?
2. To what extent do SAEBRS-TRS scores exhibit statistical homogeneity across estimates of internal consistency and relations to other variables?

3. To what extent do correlation coefficients vary as a function of moderator variables (e.g., time of administration, urbanicity, grade level)?

Study 2: SAEBRS-TRS Measurement Invariance Study

Despite the importance of equitable assessment practices, relatively little is known about frequently used universal screening tools and their functioning across a wide variety of student populations. A central purpose of assessment is to measure “true” differences among individuals, in order to allocate time and resources to those in need of support. While some of these differences may reflect characteristics we intend to measure, other differences may be due to the assessment itself (Leonardo & Grubb, 2019). Yet, equitable assessment can be a powerful mechanism for linking students to needed support and should be employed to identify and provide support for all students despite access or advantage.

Measurement invariance provides a framework for evaluating potential inequities. Specifically, measurement invariance examines the extent to which an assessment or items on an assessment perform differently across groups. Unfortunately, the examination of invariance is often abandoned until other assumptions, such as a measure’s diagnostic accuracy, are met (Pendergast et al., 2017). This diminished priority is concerning, as the invariance or equivalence of constructs and underlying psychometric properties across groups cannot be assumed (Edyburn et al., 2020). In consideration of the need for universal screening tools that function similarly across student characteristics, and the findings from Study 1, this study aims to evaluate the measurement invariance of the SAEBRS-TRS. Specifically, the results of Study 2 seek to address the following research questions:

1. To what extent do SAEBRS-TRS scores exhibit internal consistency within the study sample?

2. To what extent is concurrent validity evidence demonstrated when comparing SAEBRS-TRS scores and concurrent measures of student functioning (i.e., Office Discipline Referrals, attendance)?
3. To what extent does the SAEBRS-TRS exhibit measurement invariance when used with a variety of student characteristics (i.e., racial or ethnic identity, sex assigned at birth, eligibility for free or reduced-price lunch)?

Purpose

Continued research on the psychometric integrity and utility of screening tools across racially, ethnically, geographically, and economically diverse populations is needed. Replication of validation studies across settings with representative samples is critical for ensuring that measures can accurately and equitably identify and serve all students in need. However, with limited time and resources, it is often challenging to replicate validation studies across numerous settings and populations. In consideration of replicability challenges, systematic reviews and meta-analyses provide a feasible option for promoting the continued research on psychometrics. As such, Study 1 will evaluate the degree of extant validity evidence for the SAEBRS-TRS, including the conditions and populations under which evidence to support use exists.

Although syntheses of extant literature may assist in understanding the conditions under which a measure might be appropriately used (i.e., Study 1), findings highlight the need for independent investigators to conduct primary studies as well. In particular, the invariance or equivalence of a measure across student characteristics or populations. Therefore, Study 2 will investigate the measurement invariance of the SAEBRS-TRS across various student characteristics (i.e., racial or ethnic identity, sex assigned at birth, eligibility for free or reduced-price lunch). This study will begin to address the appropriateness of the measure's use across a variety of student characteristics; an essential element of the validation process that has yet to be examined. Together, this multi-study dissertation will inform both practice and research efforts through the psychometric evaluation of a widely used universal screening tool.

Chapter 2

Study 1: Technical Adequacy of the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale: A Systematic Review & Quantitative Synthesis

Abstract

Early identification of student social, emotional, and behavioral (SEB) concerns is vital to prevent and ameliorate future difficulties. Yet, assessment practices rely on sound validity arguments to accurately identify students who may benefit from support. As the strength of a validity argument relies on the extent of validity evidence, synthesis of extant evidence may facilitate appropriate use. Therefore, this systematic review synthesized peer-reviewed empirical studies and unpublished dissertations and theses of the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS) conducted in educational settings. Studies were included if they assessed the technical adequacy (i.e., validity evidence) of the revised version of the SAEBRS-TRS. Data were extracted and coded for sample characteristics, procedural characteristics, evidence of validity, and quality appraisal. Data were synthesized by source of validity evidence and examined for potential homogeneity. Results of the systematic search identified 29 studies meeting inclusion criteria, consisting of 65,317 students across K-12 grade levels. Overall, evidence of validity for SAEBRS-TRS scores was promising yet limited in several respects. Findings highlight limitations in the existing literature concerning the diversity of samples, a lack of evidence based on response processes and measurement bias, the relative weakness of the Emotional Behavior subscale, and the need for further exploration of the factor structure for the revised version of the measure. Future research is needed to continue examining the evidence of validity of the measure.

The prevalence of social, emotional, and behavioral (SEB) needs for children and adolescents has expanded significantly in the wake of the COVID-19 pandemic (Fitzpatrick et al., 2020). We also know that a substantial proportion of those needs are unmet, as only half of those *diagnosed* with a mental health condition subsequently receive mental health services (Whitney & Peterson, 2019). Consequently, there is a significantly high degree of unmet need. One way to better support child and adolescent SEB development, including their mental health, is through accessible prevention and intervention services, both of which necessitate accurate and early identification of needs in the SEB domains.

Educational settings are an important access point for SEB services (Duong et al., 2020) and serve as a promising setting for monitoring these issues (Lyon & Bruns, 2019). Early identification and prevention efforts for the SEB domains have naturally landed in educational settings for numerous reasons, including accessibility and feasibility (Bruhn et al., 2014). To this point, schools have the opportunity to adopt a public health approach (i.e., disease surveillance) to support the monitoring of student SEB development (Herman et al., 2012). Successful monitoring and subsequent action depend on accurate identification of student SEB concerns within a multi-tiered system of supports. As a result, schools can build capacity for addressing SEB challenges, subsequently impacting student outcomes (Miller et al., 2015).

Recognition of the current state of SEB concerns and outcomes across the nation has enhanced the need for universal screening methods in schools, to improve the early identification of students who may benefit from support (Singh et al., 2020). In light of these calls for school-based universal screening in the SEB domains, it is important to understand the strengths and weaknesses of various approaches (Miller et al., 2015). Therefore, the purpose of the present study is to systematically examine and analyze the technical adequacy of one widely used universal screening tool for the SEB domains: the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS; Kilgus et al., 2016).

The Current State of Universal Screening

Youth with social, emotional, and behavioral (SEB) difficulties are markedly underidentified in schools (Merikangas et al., 2010). The paucity of school-based SEB services and supports is alarming, especially considering more than 20% of students will display SEB needs at some point in their schooling (Merikangas et al., 2010; Perou et al., 2013). School-based universal screening in the SEB domains offers a promising option to preemptively identify students in need of support, in hopes to ameliorate the risk of future negative outcomes. This is important, as practical and accurate screening methods can inform decisions about where to focus resources for individual students, classrooms, grades, or schools demonstrating need.

In recent years, the increased awareness and prevalence of the SEB domains has resulted in a variety of approaches to identifying children and adolescents in need of support. Of these approaches, universal screening via teacher rating scales is one common method utilized. Teachers are ideally situated to identify children and adolescents who may benefit from SEB support, as they are a critical source of information regarding student functioning. To this point, teachers are well-positioned to identify children and adolescents who may benefit from SEB support as they can observe behavior across situations (e.g., academic work, social interactions). Thus, teacher-report provides essential information regarding student functioning in the school context.

Teacher ratings also have advantages above other informant methods, such as student self-report, as some students are too young to accurately self-report (Levitt et al., 2007). Additionally, some methods of identification are less predictive of outcomes or provide limited information beyond teacher ratings alone. For example, schools frequently rely on Office Discipline Referrals (ODRs) as one means of identifying students with externalizing behavior problems. Yet, these data often display low reliability due to inconsistency in teacher referrals, underrepresentation of students with internalizing problems, and disproportionate identification of students of color (Bruhn et al., 2014; Kalberg et al., 2010; Skiba et al., 2011). As such, more accurate and comprehensive identification of children and adolescents who may benefit from additional support is critical and is the desired outcome of high-quality screening via teacher-report.

Evaluation of Universal Screening Methods

The defensibility of universal screening methods rests mainly on the psychometric qualities and appropriate use of the measures. Importantly, there are necessary steps researchers must take to ensure appropriate development and validation of measures. Experts have advised that assessment research focuses on improving three main components: contextual appropriateness, usability, and technical adequacy (Glover & Albers, 2007). Although this research must support the technical properties of these novel tools, it must also be synthesized in a readily available and accessible format for a variety of audiences to consume.

Evaluation of these tools can occur by applying the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014). Guidelines from the *Standards* suggest data be extracted on five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and consequences of testing. Evidence based on test content refers to “the themes, wording, and format of the items, tasks, or questions on a test” (American Educational Research Association et al., 2014, p. 11). This evidence is typically obtained through development of items (DOI) from literature, reviews of other measures, diagnostic/eligibility criteria, and expert opinion.

According to the *Standards*, evidence based on response processes concerns “the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers” (p. 15). Evidence based on internal structure refers to the factor structure of item responses or subscales of a test. Evidence based on relations to other variables refers to the convergent, discriminant, concurrent, and predictive characteristics in comparison to other constructs and measures. Last, evidence based on consequences of testing refers to a test’s intended or unintended impact. As noted in the *Standards*, this source of validity is typically only examined for large-scale educational assessments (American Educational Research Association et al., 2014). Together, understanding the strengths and limitations of available evidence in these areas is critical for appropriate interpretation and use of a measure.

The Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS)

The SAEBRS-TRS is a brief 19-item rating scale that aims to identify students at-risk for social, academic, and emotional behavior problems. The SAEBRS-TRS is comprised of one full scale (Total Behavior [TB]) and three subscales (Social Behavior [SB], Academic Behavior [AB], Emotional Behavior [EB]). The measure is completed via paper-and-pencil or online administration. In addition to teacher-report, the SAEBRS-TRS has parent- and student-report versions. When completing the SAEBRS-TRS, teachers rate the frequency each student displayed various behaviors over the previous month. Ratings are completed using a 4-point Likert scale (0 = *never*, 1 = *sometimes*, 2 = *often*, 3 = *almost always*).

Originally, the instrument was developed to include only two subscales (SB, AB) but was subsequently revised to include three subscales (SB, AB, EB). The original version informed the theoretical interpretation and current version of the measure, which was revised to incorporate domains of mental health symptomology, including emotional behaviors (Kilgus et al., 2016). As such, the instrument developers suggest using the measure to collect information related to student functioning across social, emotional, and behavioral domains to inform intervention efforts. Considering the importance of accurate identification of students in need of intervention, it is critical to understand the body of evidence supporting the use of the SAEBRS-TRS for universal screening within an assessment-to-intervention framework.

Purpose

As experts in data-based decision-making and psychometrics within educational settings, school psychologists should assist in appropriately selecting, using, and interpreting quality assessment tools. Although available psychometrics of a universal screener are usually available in technical manuals, this information may not be the most comprehensive, accessible, or up-to-date, particularly considering the ongoing and continuous process of assessment research. Moreover, journal articles are often expensive to access and difficult to locate. Practitioners and decision-makers examining potential universal screening tools may not have the time or resources to consume research, leading to significant barriers to obtaining this information. Because of this, systematic reviews are critical for assisting both practitioners and researchers in making informed decisions about the appropriateness of the measure by having all available evidence consolidated. Therefore, the purpose of this study was to provide a critical overview and analysis of extant validity evidence of the SAEBRS-TRS. Three specific research questions were of interest:

1. What evidence of validity currently exists for the SAEBRS-TRS (e.g., response processes, test content, internal structure, relations to other variables, consequences of testing)?
2. To what extent do SAEBRS-TRS scores exhibit statistical homogeneity across estimates of internal consistency and relations to other variables?
3. To what extent do correlation coefficients vary as a function of moderator variables (e.g., time of administration, urbanicity, grade level)?

Method

Data Collection

Search Procedures

The search was conducted in November of 2021 after the study was registered through Open Science (doi: 10.17605/OSF.IO/GCA72). The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement guidelines (PRISMA; Page et al., 2021). The identification of potential studies occurred using two methods in order to perform a comprehensive search of the literature. First, an electronic search was conducted using the following databases: ERIC via EBSCO, Education Source, PsycINFO, Ovid Medline, and ProQuest Dissertations & Theses. Titles and abstracts were searched using the following descriptors: (“SAEBRS” OR “Social, Academic, and Emotional Behavior Risk Screener” OR “Social, Academic, Emotional Behavior Risk Screener”). No date restrictions were set. Titles and abstracts were reviewed for inclusion; then full text manuscripts were reviewed to determine applicability to the current literature review. Second, an ancestral search of the reference sections of included studies was conducted. Titles and abstracts of each reference were scanned to determine eligibility and studies meeting criteria were scanned at the full-text level.

Inclusion Criteria

Published studies and unpublished dissertations and theses written in English were selected for inclusion if they (a) assessed the evidence of validity of scores from the revised version of the SAEBRS-TRS; (b) involved primary data collection; (c) were conducted in a K-12 school setting; and (d) were conducted in the United States. Further, articles were excluded from data extraction if they did not administer the SAEBRS-TRS (e.g., administered parent- or student-report instead) or if they administered the original version of the measure consisting of only two subscales (SABRS). However, it should be noted that information from the original measure was included to provide context for the initial development of the measure and subsequent studies.

Data Coding

Data were extracted and coded across four categories: (a) sample characteristics; (b) procedural characteristics; (c) evidence of validity; and (d) quality appraisal. Coding for the third category (i.e., evidence of validity) included the extraction of reliability and validity coefficients consistent with the *Standards* (American Educational Research Association et al., 2014). Two graduate students independently coded 30% of included articles and inter-coder reliability (ICR) was calculated (Cooper & Hedges, 1994).

Sample Characteristics

Variables relating to the sample characteristics of each study were coded. Variables included the scope of the screening data (e.g., single school, single district), setting of the school or district (e.g., public, private), geographic location, urbanicity, grade(s) of sample, teacher characteristics (e.g., gender identity, racial or ethnic identity, years of teaching experience), student characteristics (e.g., sex assigned at birth, racial or ethnic identity), and school-wide characteristics (e.g., percentage of students eligible for special education services, percentage of students eligible for free or reduced-price lunch, percentage of students identified as Multilingual Learners).

Procedural Characteristics

Variables relating to the procedural characteristics of each study were coded. Variables included the number of administrations, time of administration (e.g., fall, winter, spring), method of administration (e.g., electronic, paper-pencil), standardized comparison measures, and unstandardized comparison measures.

Evidence of Validity

Variables relating to the evidence of validity of each study were coded. First, data related to evidence of test content was examined. As the measure development and content validation process was conducted with the original version of the SAEBRS-TRS (i.e., SABRS), these studies were not formally included in the systematic data extraction. However, research conducted using the original version provides critical information on the evidence of test content (American Educational Research Association et al., 2014). As such, a narrative review of available evidence is provided to aid in a more comprehensive understanding of the measure's development.

Next, data related to evidence based on internal structure was extracted. Variables included relevant data and estimates of reliability, factor analysis, and measurement bias. In the current framework proposed by the *Standards*, reliability is considered evidence of internal structure (American Educational Research Association et al., 2014). Data was extracted for three types of reliability evidence: internal consistency, interrater reliability, and test-retest reliability. Data related to evidence based on relations to other variables was extracted for convergent, discriminant, predictive, and concurrent validity evidence. Evidence based on response processes (e.g., response times, eye-tracking, focus groups) and consequences of testing (e.g., long-term follow up, impact on students) were not identified in any of the included articles, resulting in an empty search. Findings of the systematic review were compared to interpretive guidelines widely used in the literature, as well as in comparison to a similar universal screening tool for the SEB domains (Behavioral and Emotional Screening System, Third Edition [BESS-3]; Kamphaus & Reynolds, 2015). The BESS-3 was utilized as a comparison due to its similarity to the SAEBRS-TRS, as it is a widely used, brief, teacher-report screening tool.

Quality Appraisal

As systematic reviews are only as good as the studies used to derive the data, proper assessment of study quality is necessary to ensure high-quality review. Yet, challenges exist in the appraisal of methodological quality related to measurement studies, as study characteristics vary greatly. As no existing quality appraisal tool appropriately and comprehensively evaluates characteristics central to this study's purpose, a novel quality appraisal measure was created by adapting and integrating multiple existing measures (e.g., COnsensus-based Standards for the selection of health status Measurement INstruments [COSMIN]; Cochrane Assessing Risk of Bias in Included Studies; Guidelines for Reporting Reliability and Agreement Studies). Quality indicators included the following: replicability of the context or setting, sample, and procedures; suitability of data analysis techniques; order effects; missing data; threats to validity; and attrition. Each characteristic was given a quality rating and summed for a maximum total score of 22.

This approach is consistent with recommendations made by Cooper (2017), who suggested a mixed-criteria approach to evaluate study quality for research synthesis. Cooper

(2017) suggests combining two prominent approaches to quality appraisal: the threats-to-validity approach and the methods-description approach. By combining these into a mixed-criteria approach, as suggested prior, we are able to critically assess the methodological features of studies.

Data Analysis for Quantitative Synthesis

Transformation of Internal Consistency Estimates

To examine evidence of internal consistency, coefficient alphas and coefficient omegas were extracted. Although there are limitations of coefficient alpha as a reliability estimate (e.g., Sijtsma, 2009), it is the most commonly used estimate of internal consistency. As the distribution of coefficient alpha and coefficient omega are often negatively skewed, transformation is necessary in order for the data to be normally distributed. For reliability coefficients that range between 0 and 1 (e.g., coefficient alpha, coefficient omega), the most suitable transformation is suggested by Bonett (2002). In consideration of the possible non-normality of the data, coefficient alphas were transformed using the Bonett-transformation (Beretvas & Pastor, 2003; Borenstein et al., 2009). Following the transformation, the sampling variance of the Bonett-transformed coefficient alphas were used to determine inverse variance weights. Last, these were transformed back into their original metric for reporting.

Transformation of Concurrent Validity Estimates

To examine evidence of concurrent relationships, Pearson correlation coefficients were extracted. For validity coefficients that range between −1 and 1 (e.g., Pearson correlation coefficients), the most suitable transformation is Fisher's Z. Therefore, correlation coefficients were transformed into Fisher's Z-scores to adjust for possible non-normality of the data (Beretvas & Pastor, 2003; Borenstein et al., 2009). Following this transformation, the sampling variance of the Fisher's Z-scores were used to determine inverse variance weights. Last, these were transformed back into their original metric for reporting. As criterion measures vary greatly based on the domain measured, coefficients were disaggregated based on the following: social, emotional, and/or behavioral domains and academic domains.

Homogeneity Analysis

Homogeneity analysis, also referred to as statistical heterogeneity, compares the amount of observed variance with the amount of variance expected by sampling error alone (Cooper, 2017). A common statistic for homogeneity analysis, the *Q*-statistic, was calculated for both internal consistency coefficients and correlation coefficients identified in the systematic review. The *Q*-statistic is used in traditional meta-analysis to examine variability between and within studies attributable to systematic differences in research designs (Hedges & Olkin, 1985; Yeo, 2010). Specifically, the *Q*-statistic determines the relative homogeneity of the collective individual estimates represented by each weighted mean estimate of internal consistency and concurrent validity. While the *Q*-statistic is not an inferential test and cannot relate to other variables, it does allow for the comparability of levels between groups (Cooper, 2017). The *Q*-statistic was quantified using I^2 which examines the portion of total variance in effect sizes due to between-study variance (Cooper, 2017). A statistically significant *Q*-statistic along with I^2 values above 75% suggest considerable heterogeneity between studies (Higgins & Thompson, 2002). All analyses were conducted using the *R* package *metafor* (Viechtbauer, 2010).

Moderator Analysis

One approach to examining potential moderators of study characteristics is to calculate the homogeneity of each characteristic separately, by repeating the calculation of *Q*-statistics (Cooper, 2017). This approach allows for the identification of potentially confounding study characteristics, by comparing the variance of individual effect sizes to sampling error alone (Cooper, 2017). Moderator analyses were only conducted for correlation coefficients of the TB scale, due to an insufficient sample size for coefficient alpha and the SB, AB, and EB subscales. The moderator variables included the following: time of administration (fall, winter), urbanicity (urban, suburban, rural), and grade level (elementary only [K-5], secondary only [6-12]). A subcategory was eliminated for time of administration (spring) as only one study administered reported coefficients for this time, separate from other times of administration (i.e., winter and spring examined together). Additionally, the authors intended to examine the SEB and academic domains separately; however, this was not feasible due to the low number of coefficients. For each subcategory within each moderator variable, the weighted average effect size and 95% confidence intervals are reported. All analyses were conducted using the *R* package *metafor* (Viechtbauer, 2010).

Results

As depicted in Figure 1, the initial search yielded 148 studies. Removal of duplicates resulted in 100 unique studies. Screening of the titles and abstracts resulted in 81 studies for potential inclusion. These records were screened at the full-text level, resulting in the exclusion of 53 studies. Of the 53 excluded studies, 44 studies did not assess psychometric properties of the SAEBRS, five studies did not assess the teacher version, four studies did not include primary data collection, and one study was conducted outside of the United States. Consequently, a total of 27 studies from the electronic search met eligibility criteria and were included in the review. Additionally, two studies meeting inclusion criteria were identified through ancestral search. Of these 29 studies, 20 studies were published in peer-reviewed journals, eight studies were unpublished dissertations, and one study was an unpublished thesis. Published studies that were included in the review were retrieved from a total of 10 peer-reviewed journals. At least one instrument developer was an author on 18 of the 20 studies published in peer-reviewed journals. Studies included in this review are indicated with an asterisk in the reference section.

Sample Characteristics

Sample characteristics of studies are presented in Table 1. Across studies, student participants varied in regard to sex assigned at birth, racial or ethnic identity, eligibility for special education services, and primary language. Overall, the majority of students were male (52.6%), White (45.2%), and non-Hispanic (80.5%). Of the total sample, 26.3% of student participants were identified as eligible for special education services, 55.9% as receiving free or reduced-price lunch, and 24.1% as Multilingual Learners. Teacher characteristics were only reported for nine of 29 studies; therefore, they should be interpreted with caution. Of these studies, the majority of teachers were female (92.1%), White (82.7%), and non-Hispanic (97.1%).

Procedural Characteristics

Procedural characteristics of studies are presented in Table 2. The majority of studies (*k* = 21) examined both the full scale (TB) and all three subscales of the measure (SB, AB, EB). However, multiple studies examined the full scale only (*k* = 6) or the subscales only (*k* = 2). Most studies occurred at the elementary level (*k* = 17) or across grade levels (e.g., elementary and middle; *k* = 10). The majority of studies (*k* = 21) administered the SAEBRS-TRS on a single occurrence over the course of one academic year. The time of administration varied by study, with five studies screening in fall, four studies in winter, three studies in spring, and nine studies with multiple administrations (e.g., fall and spring). Multiple studies did not report the time of administration (*k* = 7). With regard to settings of included studies, the geographical regions of studies spanned the nation; however, the majority were conducted in the Midwest (*k* = 11). Last, the urbanicity of studies varied, with the majority conducted in urban settings (*k* = 11), followed by rural settings (*k* = 5), or suburban settings (*k* = 4). Additionally, five studies occurred across a combination of settings (e.g., urban and suburban) and four studies did not report the locale.

Systematic Review

Test Content

As noted previously, evidence based on test content was initially investigated for the original version of the measure, the SABRS. Studies examining content validity evidence for the SABRS are indicated with a double asterisk in the reference section. Kilgus et al. (2013) described the content validation and measure development process for the original SABRS. First, instrument developers discussed the purpose of the measure and the variables to be included in the measure. To develop the items, the developers attended to prior research, specifically in the following domains: social competence, academic enablers, and developmental cascade models. Next, the developers created an item pool consisting of 29 items, with nine items related to academic behavior and 20 items to social behavior (Kilgus et al., 2013). Items were then systematically reviewed by three school psychology professors and one graduate student. Reviewers placed each item into a single domain, either SB or AB, then rated their confidence in both the category and relevance of the item (Kilgus et al., 2013).

Two indices, the Content Validity Index (CVI) and Factorial Validity Index (FVI) were reviewed to determine the removal or revision of items based on expert feedback. The CVI provided an estimate of overall expert opinions regarding item representativeness. To calculate, the instrument developers divided the number of experts who rated the item as relevant by the overall number of experts. The FVI provided a measure of expert assignment of each item to the appropriate category. To calculate, the instrument developers divided the number of experts who appropriately assigned the item by the overall number of experts. Subsequently, items were removed or revised if the CVI or FVI fell below .80, the average certainty rating was equal to or less than 2.00, or multiple content experts provided specific recommendations through their open-ended feedback. Ultimately, two items were revised due to low FVI, no items were removed or revised due to low CVI, and three items were removed or revised due to low certainty. Further revisions based upon open-ended expert feedback included the removal of six items, revision of five items, and addition of one item. The resulting SABRS measure consisted of 21 items (Kilgus et al., 2013).

The instrument developers revised the measure to include the EB subscale, with the content validation process being described in von der Embse et al. (2016). Consistent with development of the SB and AB subscales, the EB subscale included items pertaining to both adaptive and maladaptive behaviors of emotional functioning. This process resulted in the generation of a 26-item pool. Similar to the original content validation process, four content experts, identified as school or counseling psychology professors, reviewed the items. First, experts reviewed a definition of the EB category, considering EB actions as the “ability to regulate internal states, adapt to change, and respond to challenging events” (von der Embse et al., 2016, p. 1268). Next, reviewers rated both the relevance and clarity of each item. Relevance ratings were calculated for each item and items were retained if their mean relevance was 4.50 or greater. Ultimately, this process resulted in seven items included in the EB subscale (von der Embse et al., 2016).

Response Processes

No information was located regarding evidence based on response processes.

Internal Structure

Internal consistency was most commonly assessed across studies reporting reliability evidence. Test-retest and interrater reliability were also reported in some cases.

Internal Consistency. Internal consistency of scores, reported as coefficient alpha, were reported for five studies. One study reported the omega coefficient, which was removed from the calculation of weighted averages. Coefficient alpha values ranged from .93 to .98 for the TB scale, .88 to .98 for the SB subscale, .92 to .98 for the AB subscale, and .77 to .98 for the EB subscale. Weighted averages were .92 for the TB scale, .93 for the SB subscale, .94 for the AB subscale, and .89 for the EB subscale. Coefficients for each study are presented in Table 3.

Interrater Reliability. Interrater reliability was reported for one study (Tanner et al., 2018). The authors reported the calculation for general education teachers and special education teachers. Spearman coefficients were .57 for special education teachers and .67 for general education teachers.

Test-Retest Reliability. Test-retest reliability, reported as Pearson correlation coefficients, was examined in three studies. Test-retest windows ranged from four weeks to three months. Correlations ranged from .80 to .91 for the TB scale, .81 to .83 for the SB subscale, .73 to .79 for the AB subscale, and .71 to .72 for the EB subscale. Weighted averages were .82 for the TB scale, .82 for the SB subscale, .77 for the AB subscale, and .72 for the EB subscale. Test-retest coefficients for each study are presented in Table 4.

Exploratory Factor Analysis (EFA). EFA was only conducted with the original version of the measure (i.e., SABRS). An initial investigation supported the SABRS model for interpretation with a bifactor structure (Kilgus et al., 2013). The instrument developers conducted two EFAs. The first EFA was developed based on Pearson product-moment correlations to examine the potential of interval-scaled data, and the second EFA was developed based on polychoric correlations to examine the potential of ordinal-scaled data. The study examined the following fit statistics: chi-square goodness-of-fit test (x^2), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA).

Next, an EFA consisting of all 21 items was conducted. Summary results of the EFA were as follows: Bartlett's test of sphericity value of .95 (p < .001), measures of sampling adequacy values ranging from .89 to .98, and factor loadings ranging from .32 to .83. The developers removed items if they met one or more of the following criteria: (a) low factor loading (≤ .50); (b) multidimensionality; (c) non-normal distribution; or (d) multicollinearity with higher loading items. The review process resulted in removal of nine items and the revised 12-item measure. A second EFA consisting of these 12 items was conducted using maximum likelihood estimation. Partial support was found for the two-factor model (x^2= 161.21, p < .01; CFI = .95; TLI = .92; RMSEA = .11) and the three-factor model (x^2= 50.39, p = .03; CFI = .99; TLI = .99; RMSEA = .05). Loadings on the first two factors within the three-factor model were similar to the initial EFA, with half of the items loading on Factor 1 (SB) and the other half on Factor 2 (AB). The authors rejected the three-factor model, stating their reasoning as low Factor 3 (TB) loadings and lower than expected internal consistency. Ultimately, the developers moved forward with the two-factor model.

Confirmatory Factor Analysis (CFA). Two studies applied traditional CFA techniques to test various models of SAEBRS-TRS scores. von der Embse et al. (2016) conducted a variety of CFAs, including a unidimensional model, correlated factors model, and bifactor model. The unidimensional model displayed poor fit to the data (x^2 = 3,242.97, p < .001; CFI = .924; RMSEA = .106), the two-factor model displayed good fit (x^2= 1,824.26, p < .001; CFI = .961;

RMSEA = .077), and the bifactor model displayed good fit and improvement compared to the previous models (x^2 = 1,120.19, p < .001; CFI = .978; RMSEA = .061). However, this bifactor model resulted in multiple Heywood cases, indicating possible model misspecification or overfitting. As the majority of Heywood cases emerged on the general factor at the teacher-level, the developers tested the full bifactor model solely at the individual-level. No further Heywood cases were identified, and no additional modifications were made to the model. The final bifactor model demonstrated good fit to the data (x^2 = 682.32, p < .001; CFI = .988; RMSEA = .045). On the general factor, 14 of 18 factor loadings displayed satisfactory fit and 16 of 18 items displayed statistically significant loadings (p < .001). The two items that did not have significant loadings on the general factor displayed significant loadings on the specific factors.

Additionally, Kilgus et al. (2018a) tested a variety of models using CFA techniques, including a unidimensional model, two-factor model, and bifactor model. The unidimensional model displayed poor fit to the data (RMSEA = .11; CFI = .88; TLI = .87), the two-factor model displayed good fit (RMSEA = .089; CFI = .92; TLI = .91), and the bifactor model displayed good fit and improvement compared to the previous models (RMSEA = .060; CFI = .97; TLI = .96). Results of each CFA and EFA are presented in Table 5.

Measurement Bias. One study examined measurement bias of the EB subscale through differential item functioning (DIF). Izumi (2020) used DIF to evaluate the extent of measurement invariance across student characteristics for the EB subscale. To assess levels of DIF, visual methods and effect size estimates were calculated across student characteristics based on racial or ethnic identity, as well as the interaction between racial or ethnic identity and sex assigned at birth. Effect sizes were calculated at the item-level (expected score standardized difference;

ESSD) and test-level (expected test score standardized difference; ETSSD). These effect sizes were interpreted by the authors as follows: small > .20; medium > .50; large > .80.

In accordance with the author's interpretive guidelines, results indicated the smallest DIF for students of two or more races and largest DIF for Black students. For Black students, item-level effect sizes (ESSDs) ranged from -.11 to -.87, with a test-level effect size (ETSSD) of -.56. For Hispanic students, item-level effect sizes (ESSDs) ranged from -.05 to .25, with a test-level effect size (ETSSD) of .11. For students of two or more races, item-level effect sizes (ESSDs) ranged from -.07 to .18, with a test-level effect size (ETSSD) of .05. For White students, item-level effect sizes (ESSDs) ranged from -.01 to .54, with a test-level effect size (ETSSD) of .33. Overall, the effects of DIF at the test-level were negligible to small for students of two or more races and Hispanic students, and moderate for Black students and White students. Additionally, the study examined the interaction between racial or ethnic identity and sex assigned at birth via DIF analyses. Effect sizes were generally larger for males compared to females. For each racial or ethnic identity, levels of effect size (i.e., small, medium, large) did not change when disaggregated by sex assigned at birth. It is unclear the extent to which these findings hold for the SB subscale, AB subscale, and TB scale.

An additional study was located for the original version of the measure, the SABRS (Pendergast et al., 2017). Analyses were performed for invariance across racial or ethnic identity (White and Black students). Specifically, the author conducted a confirmatory factor analysis using weighted least squares mean and variance adjusted (WLSMV) estimation with a two-factor model (SB and AB). The results supported configural, metric, and scalar invariance, with fit indices ranging from borderline to strong across groups. Specifically, the authors proposed the findings for the configural model as follows: the Confirmatory Fit Index (CFI) of .993 was strong and the Root Mean Square Error of Approximation (RMSEA) of .077 was borderline. For the metric model, authors identified the CFI of .994 as strong, RMSEA of .069 as borderline, and change-in-model fit indices as strong (non-significant Δx^2; ΔCFI < .01; ΔRMSEA < .015). For the scalar model, authors noted the CFI of .994 as strong, RMSEA of .062 was borderline, and change-in-model fit indices were strong (non-significant Δx^2; ΔCFI < .01; ΔRMSEA < .015). Overall, findings provided preliminary support for the equivalence across these groups for the SABRS; however, it is unclear the extent to which these findings hold for the revised measure (SAEBRS). Additionally, analyses were conducted with a two-factor model, despite prior support for a bifactor model for the SABRS.

Relations to Other Variables

Measures of Academic Outcomes. Ten studies examined relations to other variables for academic outcomes. Absolute correlation coefficients ranged from .00 to .95 for the TB scale, .01 to .91 for the SB subscale, .01 to .89 for the AB subscale, and .01 to .85 for the EB subscale. Overall weighted averages, calculated using absolute values, were as follows: .25 for the TB scale, .12 for the SB subscale, .33 for the AB subscale, and .15 for the EB subscale. When disaggregated by academic domain, coefficients of reading measures ranged from .04 to .45, writing measures ranged from .03 to .34, and math measures ranged from .25 to .46 for the full scale. Weighted averages for the TB scale by academic domain were as follows: .24 for reading, .19 for writing, and .39 for math. Coefficients for each study are presented in Table 6.

Measures of SEB Outcomes. Nine studies examined relations to other variables for SEB outcomes. Absolute correlation coefficients ranged from .24 to .79 for the TB scale, .28 to .61 for the SB subscale, .18 to .83 for the AB subscale, and .06 to .62 for the EB subscale. Overall weighted averages, calculated using absolute values, were as follows: .58 for the TB scale, .53

for the SB subscale, .50 for the AB subscale, and .45 for the EB subscale. Coefficients for each study are presented in Table 7.

Classification/Diagnostic Accuracy. Eight studies examined evidence based on classification/diagnostic accuracy. Given the sample variability across studies, cut scores were reported in various ways. Some studies reported a single cut score on the basis of what the study authors considered optimal for the sample, whereas others reported multiple cut scores given specific parameters. Across the eight studies examining classification/diagnostic accuracy, optimal cut scores ranged from 27 to 42 for the TB scale, 6 to 15 for the SB subscale, 6 to 15 for the AB subscale, and 12 to 18 for the EB subscale.

Additional findings for the TB scale included Area Under the Curve (AUC) values ranging from .65 to .99, sensitivity from .39 to .97, specificity from .77 to .95, positive predictive values from .19 to .74, negative predictive values from .90 to .99, positive likelihood ratios from 5.63 to 13.86, and negative likelihood ratios from .03 to .14. For the SB subscale, AUC values ranged from .85 to .97, sensitivity from .80 to .93, specificity from .71 to .93, positive predictive values from .22 to .72, negative predictive values from .94 to .99, positive likelihood ratios from 2.79-7.90, and negative likelihood ratios from .08 to .27. For the AB subscale, AUC values ranged from .85 to .96, sensitivity from .76 to .93, specificity from .73 to .91, positive predictive values from .22 to .62, negative predictive values from .92 to .99, positive likelihood ratios from 2.85 to 6.64, and negative likelihood ratios from .08 to .32. Lastly, for the EB subscale, AUC values ranged from .84 to .97, sensitivity from .60 to .90, specificity from .65 to .91, positive predictive values from .26 to .55, negative predictive values from .89 to .99, positive likelihood ratios from 2.57 to 6.43, and negative likelihood ratios from .12 to .32. Findings are presented in Tables 8-11.

Consequences of Testing

No information was located regarding evidence based on the consequences of testing.

Quantitative Synthesis

Homogeneity Analysis

Homogeneity analyses were conducted for coefficients of internal consistency and correlation coefficients. Although analyses were performed using different transformations for raw estimates, tables display the estimates once back transformed to the corresponding coefficient metric. Table 12 includes weighted estimates with 95% confidence intervals, *Q*-statistics, and I^2 values expressed as a percentage. For internal consistency, evidence of heterogeneity was found for all scales. The *Q*-statistics were statistically significant ($p < .05$) and I^2 values indicated considerable heterogeneity for all three subscales, ranging from 99.43% to 99.71%, and moderate heterogeneity for the full scale (62.85%). For correlation coefficients of academic measures, evidence of heterogeneity was found for most scales. The *Q*-statistics were significant for the TB scale, SB subscale, and AB subscale ($p < .05$), and I^2statistics indicated considerable heterogeneity for the TB scale (98.67%) and AB subscale (88.76%), moderate heterogeneity for the full scale (52.62%), and mild heterogeneity for the EB subscale (15.38%). Lastly, for correlation coefficients of SEB measures, evidence of heterogeneity was found for all scales. The *Q*-statistics were significant for the TB scale, SB subscale, and AB subscale ($p < .05$), and I^2 statistics indicated considerable heterogeneity for the TB scale (99.45%), SB subscale (98.45%), AB subscale (98.78%), and EB subscale (98.01%). The *Q*-statistic was not statistically significant for the EB subscale.

Moderator Analysis

Further analysis of homogeneity was conducted, in order to examine potential moderator variables. Table 13 displays weighted estimates with 95% confidence intervals, *Q*-statistics, and I^2 values expressed as a percentage. For the moderator variable of time of administration, evidence of heterogeneity was identified for fall and winter. The *Q*-statistics were significant ($p < .05$) and I^2 values indicated considerable heterogeneity, ranging from 99.73% for fall administration to 90.12% for winter administration. For the moderator variable of urbanicity, evidence of heterogeneity was found for urban, suburban, and rural samples. The *Q*-statistics were statistically significant ($p < .05$) and I^2 values indicated considerable heterogeneity for all three subscales, ranging from 99.33% for urban populations to 99.56% for rural populations. Lastly, the moderator variable of grade level showed evidence of heterogeneity for both elementary and secondary grade levels. The *Q*-statistics were statistically significant ($p < .05$) and I^2 values displayed considerable heterogeneity, ranging from 99.31% for elementary grades to 99.73% for secondary grades.

Quality Appraisal

Study quality was assessed using an author-created quality appraisal measure. An overall rating of quality was given for each study as well as ratings for each quality indicator, including the following: replicability of the context or setting, sample, and procedures; suitability of data analysis techniques; order effects; missing data; threats to validity; and attrition. Each quality indicator was operationally defined, and generally followed the qualitative rating system of *Not Applicable, 0 (did not address the quality indicator), 1 (somewhat addressed the quality indicator),* or *2 (fully addressed the quality indicator)*. Quality appraisal scores of the methodological characteristics of included studies ranged from 9 to 19 (mean = 12.4), out of a maximum possible score ranging from 14 to 22 (mean = 18.9). Consequently, the quality of studies included were variable, but fair overall. However, there is room for improvement related to study quality for future studies related to the SAEBRS-TRS, as a lack of methodological quality diminishes the ability to draw inferences from study results.

The most frequent missing information, when applicable to the study, resulting in decreased study quality included the following: procedures to minimize potential order effects at the student-level (e.g., student randomization), procedures to minimize potential order effects when multiple measures were used (e.g., counterbalancing), and adjustment for inflated Type 1 error when multiple measures were used on the same sample. Conversely, the majority of studies (greater than 50% for each characteristic) provided sufficient information to reasonably replicate the study procedure and setting, reported or addressed missing data, and provided data for attrition rates when applicable.

Discussion

Universal screening is an integral component of a multi-tiered system of supports (Glover & Albers, 2007), which allows educators to engage in data-based decision-making and early identification of students who may benefit from SEB support. Research has demonstrated universal screeners are more efficient and effective compared to commonly used methods, such as teacher nomination or Office Discipline Referrals (Dowdy et al., 2013). Given these findings and the push for schools to adopt evidence-based practices for assessment and intervention, the use of brief behavior rating scales as universal screeners has grown in popularity in recent years. The SAEBRS-TRS is one universal screening tool used to systematically assess students who may benefit from SEB support.

Researchers have yet to quantitatively synthesize the SAEBRS-TRS literature to determine the present status of the measure’s psychometric defensibility. Therefore, the purpose of this study was to synthesize empirical research related to the technical adequacy of the SAEBRS-TRS. Specific research purposes were threefold: (a) examine the existing evidence of validity by applying the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014); (b) examine statistical heterogeneity across estimates of internal consistency and concurrent validity; and (c) examine the extent to which correlation coefficients vary as a function of moderator variables. This review furthers our understanding of the technical adequacy of the SAEBRS-TRS. Findings suggest limitations in the literature, specifically related to the geographic, racial, and ethnic diversity of samples, a lack of evidence based on response process and consequences of testing, the relative weakness of the EB subscale, and the need for further exploration of factor structure related to the revised version of the measure.

The first purpose of the study was to synthesize evidence of validity in the following domains: response processes, test content, internal structure, relations to other variables, and consequences of testing. Of note, there is currently no evidence of validity based on response processes or consequences of testing. Evidence based on response processes are critical to examine the relationship between construct fit and the response engaged in by test takers (American Educational Research Association et al., 2014). This evidence of validity may be gathered by examining qualitative responses of raters to identify how they arrived at answers or approached questions. As response processes can provide evidence of validity for various aspects of construct validity and explanations of scores, it is recommended future research examine this facet of validity evidence for the SAEBRS-TRS.

Additionally, no evidence was located relating to the consequences of testing. As noted in the *Standards for Educational and Psychological Testing*, "Some consequences of test use follow directly from the interpretation of test scores for uses intended by the test developer. The validation process involves gathering evidence to evaluate the soundness of these proposed interpretations for their intended uses" (American Educational Research Association et al., 2014, p. 19). The consequences of a test score can be linked to a flaw in the conceptualization of test interpretation (e.g., construct underrepresentation, construct irrelevant variance). Therefore, it is critical to synthesize the available evidence of validity in accordance with these guidelines to better understand a scale's potential boundaries in its application across populations and test settings (American Educational Research Association et al., 2014). Yet, this is an often-overlooked source of validity, as noted in the present review.

In the domain of test content, due to revision of the scale, the instrument developers first reported the content validation process for the SB and AB subscales. The content validation process for the EB subscale occurred separately; therefore, the independent content validation processes may have impacted these results. As findings related to the EB subscale were relatively weak across sources of validity evidence in comparison to the full scale and other two subscales, the measure development process warrants concern. Additionally, important aspects of measure development, such as the Content Validity Index (CVI) and Factorial Validity Index (FVI) were not reported. Specifically, the CVI provides an estimate of overall expert opinions regarding item representativeness and the FVI examines the extent that experts assigned each item to a corresponding category determined by the measure developers. While these indices were calculated for the original measure (i.e., SABRS), it is unclear if these were examined for the revised measure, as we were unable to locate information related to content development.

Similarly, another widely used screening measure, the BESS-3 Teacher, was revised from its second edition. The BESS-2 Teacher consisted of 27 items, whereas the BESS-3 Teacher consists of 20 items (Kamphaus & Reynolds, 2015). It is unclear whether a comprehensive content validation process was conducted for the revised BESS-3 Teacher, as we were unable to locate this information in the peer-reviewed literature. This lack of reporting regarding important aspects of measurement development is concerning, especially when examining the validation of a measure.

With regard to internal structure, results were supportive of the internal consistency of SAEBRS-TRS scores, with a weighted estimate of .92 for the full scale. This estimate exceeded the common threshold for low-stakes decisions (.70; Cortina, 1993). Evidence of test-retest reliability provides support for the stability of SAEBRS-TRS scores across time. Despite these promising findings, results also determined gaps in terms of evidence of interrater reliability. Only one study examined interrater reliability and the estimates were lower than desired (.57 for special education teachers and .67 for general education teachers); therefore, the consistency of scores across raters is unclear. Although the majority of interrater reliability studies typically examine consistency across different informants (e.g., teacher and parent versions of a measure), it is also important to examine consistency within settings across raters (e.g., multiple teachers). Moreover, given the existing teacher and parent versions of the SAEBRS (mySAEBRS and SAEBRS-PRS), interrater reliability is particularly important to examine consistency across raters.

Lower reliability coefficients were reported for the EB subscale across different types of reliability evidence (i.e., internal consistency, test-retest). Despite the promising findings in the domain of internal consistency, potential flaws with the methodological and analytic techniques were identified. It was noted that a single study utilized coefficient omega, and the majority examined internal consistency using coefficient alpha. This is concerning, as data represented by a bifactor model are more appropriately measured using coefficient omega (Flora, 2020). Specifically, hierarchical omega is preferable, as it represents the total-score variance due to a general construct, while also considering the multidimensional nature of the data (Rodriguez et al., 2016).

Another common technique used to examine internal structure, factor analysis, is a critical component of validity evidence. In this review, although the results were supportive of a bifactor model, issues with the analytical approaches were identified. For example, multiple analyses used maximum likelihood (ML) estimation rather than WLSMV estimation. This is problematic considering the SAEBRS-TRS consists of ordinal variables, rather than continuous variables. Continuous variables are appropriate for use with ML estimation, as the CFA model is fit to the observed Pearson product-moment covariance or correlations. Comparatively, it is more appropriate to use WLSMV with ordinal variables, as the CFA model is fit to polychoric correlations to account for the categorical nature (Finney & DiStefano, 2013). The authors defended this analytic decision by examining normality using both ML and WLSMV estimation. The authors compared findings through a sensitivity analysis, which indicated equivalence across estimation types, leading them to move forward with ML estimation. Lastly, interpretation of the model structure is questioned, considering the interpretive limitations of the bifactor model. As the bifactor model assumes general and specific factors are orthogonal (i.e., uncorrelated), this does not translate to the theoretical basis of the model, as one would assume social behavior, academic behavior, emotional behavior, and total behavior are correlated. Despite lower model fit indices, the second-order model may be advantageous due to the interpretability of the findings (i.e., oblique/correlated factors; Rios & Wells, 2014).

Further, we were unable to locate studies in regard to measurement bias. Although a scoping search for the original version of the measure identified one study examining measurement bias, analyses were only tested for invariance across age, sex assigned at birth, and eligibility for special education services. Findings provided preliminary support for the equivalence across these groups; however, it is unclear the extent to which findings are supported with the revised measure. Similar to the SAEBRS-TRS, the BESS-3 Teacher has examined potential bias when comparing across race, ethnicity, sex assigned at birth, and language (Houri, 2020; Splett et al., 2020), but we were unable to locate any studies that have specifically examined measurement invariance.

The lack of bias evaluation in this regard is highly problematic, yet not surprising, as measurement bias is often overlooked in the validation process. As such, educators must be cautious in interpreting SAEBRS-TRS scores obtained from marginalized groups, as measurement bias has not been investigated at this time. As we confront a national reckoning on issues of racial injustice and inequity, it is more important than ever that we critically evaluate potentially discriminatory practices. Assessment is often used as a means to compare students of color or students below the poverty line to more advantaged students, subsequently marginalizing them even more. Yet, nondiscriminatory assessment can be a powerful mechanism for linking students to needed supports. As schools consider the adoption of various screening tools, attention to these aspects of measurement bias is warranted. Future research remains necessary to better inform practice in this area.

Next, evidence was synthesized based on relations to other variables. Collectively, results suggest the SAEBRS-TRS is an adequate indicator of SEB and academic outcomes, with weighted estimates exceeding common thresholds. This indicates the SAEBRS-TRS theoretical basis in relation to similar constructs is strong, and the relationship with various measures of student behavior is well-established. However, it should be noted the majority of studies utilized Pearson correlation coefficients. Although Pearson correlation coefficients are appropriate in many cases, it is generally recommended that Spearman correlation coefficients be used when measuring the association of ordinal data, such as data on a Likert scale. These findings are similar to those of the BESS-3 Teacher (Kamphaus & Reynolds, 2015), which also has a strong body of evidence regarding relations to other variables, specifically of concurrent and predictive validity of academic, SEB, and school climate outcomes (Naser & Dever, 2020; Splett et al., 2020). However, the aforementioned concerns regarding analytic choices are present with this measure and others, with many reporting Pearson correlations despite the ordinal nature of the variables.

Lastly, homogeneity and moderator analyses were carried out to examine variability across internal consistency coefficients and correlation coefficients. Although coefficient alphas included in the review were generally high, considerable heterogeneity was found between included studies. The results indicated approximately 90% of the variability of coefficient alphas can be explained by the heterogeneity of true score reliabilities, with the exception of coefficient alphas for the TB scale. Further, considerable heterogeneity was identified for correlation coefficients reported in the review. Interestingly, mild heterogeneity was found for the EB subscale. Results for SEB and academic relations were varied, with greater heterogeneity identified across measures of the SEB domains. This is surprising, considering the main purpose of the SAEBRS-TRS is to assess the social, emotional, and behavioral domains of student functioning. Finally, we examined whether the internal consistency or correlational findings were moderated by time of administration, urbanicity, and grade level. Evidence of heterogeneity was found across all potential moderator variables; however, this was anticipated given the considerable heterogeneity identified prior to moderator analyses. Additionally, these findings should be interpreted with caution given the small sample size for each analysis.

Considering the range of settings and samples the SAEBRS-TRS is utilized, it would be beneficial to conduct further homogeneity analyses and moderator analyses to better capture the potential generalizability of findings. Given the small sample size included in the heterogeneity analysis, we were unable to conduct comprehensive moderator analyses inclusive of coefficient alpha, disaggregated secondary grade levels (e.g., 6-8 and 9-12), or all times of administration (e.g., spring). Despite the need for homogeneity and moderator analyses with larger sample sizes, the presence of heterogeneity in the included studies warrants caution for the generalizability of the SAEBRS-TRS score reliability and concurrent relations to other variables. A portion of the variability may be due to methodological characteristics; yet, the variability is evident and must be considered when interpreting the generalizability of scores.

Taken together, the current findings are promising yet preliminary concerning the SAEBRS-TRS technical adequacy. Findings indicate the potential of the items to predict a general factor related to SEB functioning, given the relations to other variables and measures of the SEB domains. Specifically, findings suggest the SAEBRS-TRS is more highly related to variables theoretically aligned with the SAEBRS-TRS (e.g., behavior) and more weakly related to variables less theoretically aligned (e.g., academic). Moreover, it should be noted that results at the EB subscale level are weaker in comparison to the results of the TB scale, SB subscale, and AB subscale. Therefore, caution is warranted when using and interpreting the results of the EB subscale. At this point in the development of the measure’s validity argument, educators should refrain from interpreting the scores of the EB subscale separately from the TB scale.

Limitations & Future Directions

Although this study aimed to conduct a rigorous review of the literature, we wish to note several limitations of our review that must be considered in interpreting findings. First, the study was limited by the search parameters, generating the potential for missed studies. As grey literature was minimally examined through a single database (i.e., ProQuest Dissertations & Theses), it is possible that relevant results were not captured in the search. Due to the potential for missed studies and limited inclusion of grey literature, the review was based on a relatively small number of studies. It is recommended future studies include a greater variety of unpublished literature for a comprehensive representation of evidence. Second, we were unable to locate quality appraisal measures that aligned with the purposes of this study. Therefore, study quality was examined with an author-developed quality appraisal tool. While overall quality ratings are reported, indicators are not of equal weight and should be interpreted with caution.

Third, heterogeneity across studies must be taken into consideration when interpreting the quantitative synthesis of the literature presented in this study. Although some argue the inevitably of statistical heterogeneity (e.g., Higgins et al., 2003), the statistical variation across studies in this review does not allow for generalizable conclusions. Although this limitation is common across reviews of psychometric properties due to the number of analytical options, it does not allow for thorough reporting of the state of the literature. Moreover, a single research group conducted the majority of studies. It is recommended that independent researchers examine the psychometric properties of the SAEBRS-TRS. Further, the included literature was largely conducted at the elementary level. Therefore, the results of this review should be interpreted with caution when considering use at the secondary level. Future research should prioritize the examination of the SAEBRS-TRS with older age groups, and report findings through the disaggregation of secondary grade levels (e.g., 6-8 and 9-12). Lastly, few studies reported teacher characteristics, limiting the extent to which findings may be generalized. Future research must provide adequate descriptions of all participant characteristics, including teacher characteristics, particularly in the examination of response processes.

Conclusion

The current findings provide encouraging but limited evidence of validity of SAEBRS-TRS scores. Broadly, this review has shown that evidence for the internal structure and relations to other variables is adequate. However, there is no evidence related to response processes or consequences of testing, and there is a lack of evidence for interrater reliability and measurement bias. Although evidence of the internal structure is promising, confirmatory factor analyses of the revised measure are based on exploratory results of the original measure. These issues are compounded considering the majority of SAEBRS-TRS research has been conducted by the instrument developers. Overall, this lack of evidence is concerning given the growing implementation of the measure across the nation. Researchers must prioritize the further development of the measure's validity argument, and practitioners must interpret SAEBRS-TRS scores in light of the potential limitations. Future research is required to further examine the evidence of validity of the measure.

Figure 1

Study Inclusion Process

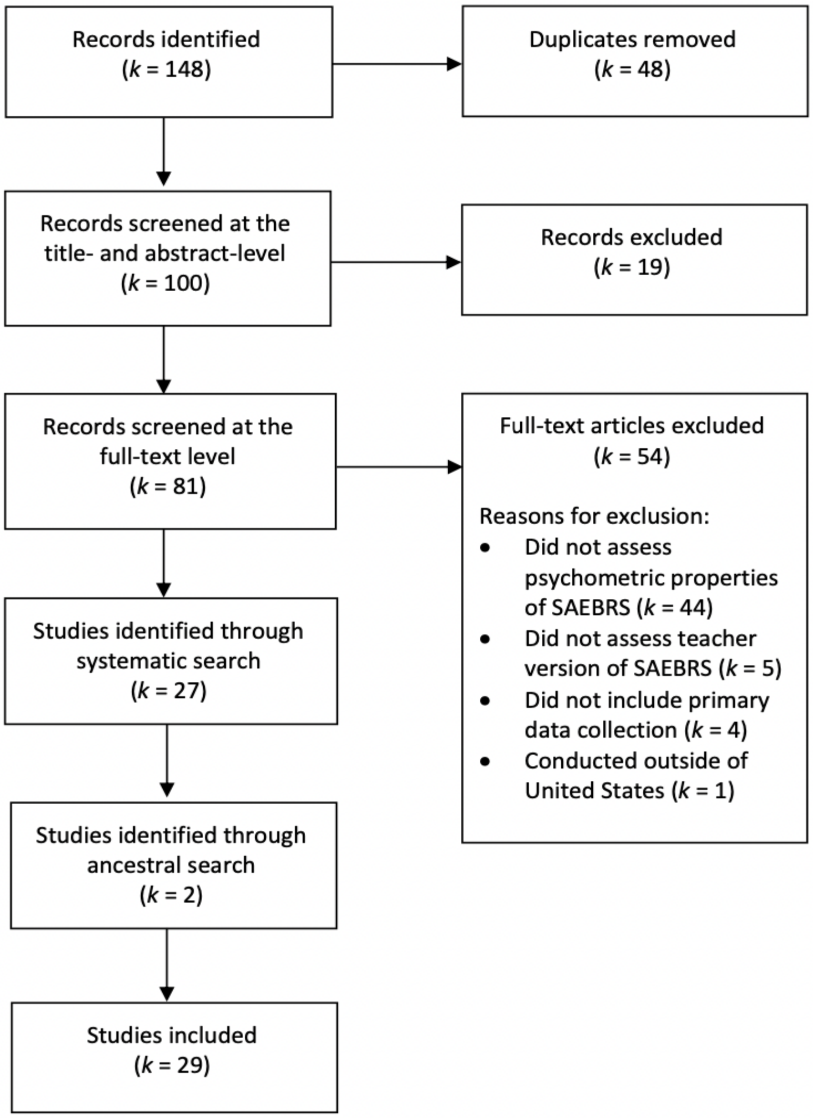

Records identified
(*k* = 148)
Duplicates removed
(*k* = 48)
Records screened at the title- and abstract-level
(*k* = 100)
Records excluded
(*k* = 19)
Records screened at the full-text level
(*k* = 81)
Full-text articles excluded
(*k* = 54)
Reasons for exclusion:
• Did not assess psychometric properties of SAEBRS (*k* = 44)
• Did not assess teacher version of SAEBRS (*k* = 5)
• Did not include primary data collection (*k* = 4)
• Conducted outside of United States (*k* = 1)
Studies identified through systematic search
(*k* = 27)
Studies identified through ancestral search
(*k* = 2)
Studies included
(*k* = 29)

Table 1

Sample Characteristics of Studies Included in Review

Characteristic	Student[a] (%)	Teacher[b] (%)
Sex Assigned at Birth		
Female	47.0	92.1
Male	52.6	7.9
Racial Identity		
White	45.2	82.7
Black or African American	26.4	11.3
Asian	2.6	5.1
American Indian or Alaska Native	1.6	1.1
Native Hawaiian or Other Pacific Islander	0.5	0
Multiple Races	6.4	4.6
Other	0.3	0
Ethnic Identity		
Hispanic	19.5	2.8
Non-Hispanic	80.5	97.1
Special Education	26.3	-
Free or Reduced-Price Lunch	55.9	-
Multilingual Learner	24.1	-

Note. Student characteristics reported for 28 studies. Teacher characteristics reported for nine studies.

[a]*n* = 65,317. [b]*n* = 973.

Table 2

Procedural Characteristics of Studies Included in Review

Characteristic	Studies, *k* (%)
Domains Assessed	
Full Scale Only	6 (21)
Subscale(s) Only	2 (7)
Full Scale & Subscales	21 (72)
Grade Level	
Elementary (K-5)	17 (59)
Middle (6-8)	2 (7)
High (9-12)	-
Multiple	10 (34)
Unknown	-
Scope	
Single School	6 (21)
Multiple Schools	19 (66)
National	3 (10)
Unknown	1 (3)
Setting	
Public	7 (24)
Private	-
National	3 (10)
Unknown	19 (66)
Geography	
Midwest	11 (38)
Northeast	4 (14)
South	2 (7)
West	1 (3)
Multiple	9 (31)
Unknown	2 (7)
Urbanicity	
Urban	11 (38)
Suburban	4 (14)
Rural	5 (17)
Multiple	5 (17)
Unknown	4 (14)
Number of Administrations	
One	21 (72)
Two	5 (17)
Three	2 (7)
Unknown	1 (3)
Time of Administration	
Fall	5 (17)
Winter	4 (14)
Spring	3 (10)
Multiple	10 (35)
Unknown	7 (24)

Table 3

Summary Findings from Studies Examining Internal Consistency

Study	*n*	Statistic	TB	SB	AB	EB
Eklund et al., 2017 (K-5)	1,044	Alpha	.94	.91	.93	.86
Eklund et al., 2017 (6-8)	1,044	Alpha	.94	.98	.94	.81
Iaccarino et al., 2017 (Fall)	1,257	Alpha	-	.96	.96	.95
Iaccarino et al., 2017 (Winter)	1,257	Alpha	-	.97	.98	.97
Iaccarino et al., 2017 (Spring)	1,257	Alpha	-	.98	.98	.98
Kilgus et al., 2016a (Study 1)	864	Alpha	.93	.89	.92	.83
Kilgus et al., 2016a (Study 2)	864	Alpha	.94	.93	.92	.77
Kilgus et al., 2016a (Study 1)	1,534	Alpha	.94	.89	.92	.82
Kilgus et al., 2016a (Study 2)	1,534	Alpha	.93	.88	.93	.79
Kilgus et al., 2016b	346	Alpha	.94	.90	.92	.86
Kilgus et al., 2018a	1,243	Omega	.98	.97	.97	.96
Kilgus et al., 2018c	1,242	Alpha	.94	.92	.93	.84
Kim & von der Embse, 2021	24,094	Alpha	-	-	.92	-
Weighted Average			**.92**	**.93**	**.94**	**.89**

Note. TB = Total Behavior scale. SB = Social Behavior subscale. AB = Academic Behavior subscale. EB = Emotional Behavior subscale. Omega coefficient removed for weighted average calculation.

Table 4

Summary Findings from Studies Examining Test-Retest Reliability

Study	*n*	Test Window	TB	SB	AB	EB
Roberson, 2019	332	3 months	.82	.83	.79	.72
Tanner et al., 2018 (Special Education)	82	3 months	.85	-	-	-
Tanner et al., 2018 (General Education)	82	3 months	.91	-	-	-
Whitley et al., 2019	175	4 weeks	.80	.81	.73	.71
Weighted Average			**.82**	**.82**	**.77**	**.72**

Note. TB = Total Behavior scale. SB = Social Behavior subscale. AB = Academic Behavior subscale. EB = Emotional Behavior. All statistics reported as Pearson correlation coefficients.

Table 5

Summary Findings from Studies Examining Factor Analysis

Measure & Study	Analysis	Model	Estimation	χ^2	RMSEA	CFI	TLI	SRMR
SABRS								
Kilgus et al., 2013	EFA	Two-factor	ML	161.21*	.11	.95	.92	-
Kilgus et al., 2013	EFA	Three-factor	ML	50.39*	.05	.99	.99	-
Kilgus et al., 2015	CFA	Unidimensional	ML	1299.497*	.217	.728	.668	.119
Kilgus et al., 2015	CFA	Two-factor	ML	527.316*	.135	.897	.871	.085
Kilgus et al., 2015	CFA	Bifactor 1	ML	203.432*	.089	.965	.945	.044
Kilgus et al., 2015	CFA	Bifactor 2	ML	129.528*	.068	.980	.968	.050
Pendergast et al., 2017	CFA	Two-factor 1	WLSMV	120.970*	.085	.993	-	-
Pendergast et al., 2017	CFA	Two-factor 2	WLSMV	108.534*	.070	.993	-	-
SAEBRS								
Kilgus et al., 2018a	CFA	Unidimensional	WLSMV	-	.11	.88	.87	-
Kilgus et al., 2018a	CFA	Two-factor	WLSMV	-	.089	.92	.91	-
Kilgus et al., 2018a	CFA	Bifactor	WLSMV	-	.060	.97	.96	-
von der Embse et al., 2016	CFA	Unidimensional	WLSMV	3242.97*	.106	.924	-	.157
von der Embse et al., 2016	CFA	Two-factor	WLSMV	1824.26*	.077	.961	-	.111
von der Embse et al., 2016	CFA	Bifactor	WLSMV	682.32*	.045	.988	-	.052

Note. χ^2 = Chi-square goodness-of-fit test. TLI = Tucker-Lewis Index. CFI = Comparative Fit Index. RMSEA = Root Mean Square Error of Approximation. SRMR = Standardized Root Mean Square Residual. EFA = Exploratory Factor Analysis. CFA = Confirmatory Factor Analysis. ML = Maximum Likelihood. WLSMV = Weighted Least Squares Mean and Variance Adjusted. Bifactor 1 = No covariance estimated. Bifactor 2 = Covariance estimated between two sets of item residuals. Two-factor 1 = Black participants. Two-factor 2 = White participants.

*p < .05

Table 6

Summary Findings from Studies Examining Relations to Other Variables (Academic Measures)

Study	*n*	Measure	TB	SB	AB	EB
Hamsho, 2020	147	WIAT-III Essay Composition	.34	.21	.36	.35
Hamsho, 2020	147	CBM-Written Expression	.31	.20	.37	.27
von der Embse et al., 2018 (Time 1)	1,158	Writing Composite	.06	.03	.15	-.07
von der Embse et al., 2018 (Time 2)	1,158	Writing Composite	.03	-.05	.16	-.04
von der Embse et al., 2018 (Time 1)	1,158	Fountas & Pinnell	.12	.07	.21	.01
von der Embse et al., 2018 (Time 2)	1,158	Fountas & Pinnell	.11	.03	.21	.02
Kilgus et al., 2017	1,058	aReading	.37	.24	.45	.24
Eklund et al., 2017	1,044	Reading: Overall	.33	.16	.40	.28
Eklund et al., 2017	1,044	Reading: Within-Group	.35	.16	.41	.30
Eklund et al., 2017	1,044	Reading: Between-Group	.24	.11	.35	.17
Roberson, 2019 (Time 1)	371	Reading Achievement	.41	.21	.55	.26
Roberson, 2019 (Time 2)	332	Reading Achievement	.45	.19	.60	.32
von der Embse et al., 2018 (Time 1)	1,158	Reading Composite	.23	.14	.40	.05
von der Embse et al., 2018 (Time 2)	1,158	Reading Composite	.15	.07	.30	.01
von der Embse et al., 2018 (Time 1)	1,158	Comprehension Composite	.19	.07	.32	.10
von der Embse et al., 2018 (Time 2)	1,158	Comprehension Composite	.29	.14	.41	.20
Kilgus et al., 2016b	346	Reading-CBM	.41	.31	.48	.28
Pendergast et al., 2018	1,461	Reading-CBM	.32	-	-	-
Whitley et al., 2019 (Time 1 BOY)	175	Oral Reading Fluency	.08	.02	.24	-.10
Whitley et al., 2019 (Time 2 BOY)	175	Oral Reading Fluency	-.04	-.03	.13	-.20
Whitley et al., 2019 (Time 1 EOY)	175	Oral Reading Fluency	.15	.04	.28	.01
Whitley et al., 2019 (Time 2 EOY)	175	Oral Reading Fluency	.09	.04	.24	-.08
Kilgus et al., 2017	1,058	aMath	.36	.20	.42	.21
Roberson, 2019 (Time 1)	371	Math Achievement	.46	.24	.59	.30
Roberson, 2019 (Time 2)	332	Math Achievement	.46	.19	.62	.34
von der Embse et al., 2018 (Time 1)	1,158	Math Composite	.25	.10	.38	.17
von der Embse et al., 2018 (Time 2)	1,158	Math Composite	.30	.14	.39	.22
Kilgus et al., 2016b	346	MAP-Math	.45	.30	.52	.32
Kilgus et al., 2016b	346	MAP-Communication	.45	.34	.52	.28
Kilgus et al., 2017	1,058	Academic Composite	.39	.20	.49	.27
Youkhanna, 2021	85	WJ-Reading Cluster	-	.08	.21	.00
Youkhanna, 2021	85	CREVT	-	-.02	.16	.24
Youkhanna, 2021	85	Vocabulary Gains	-	-.12	.00	.00
Weaver, 2019	35	Progressive Ratio Arbitrary	-.05	-.08	-.08	-.02
Weaver, 2019	35	Progressive Ratio Academic	-.21	-.25	-.03	-.23
Weaver, 2019	35	Delay Discounting Real Prizes	-.14	-.09	-.29	.06
Weaver, 2019	35	Delay Discounting Money	-.20	-.01	-.46	-.03
Weaver, 2019	35	Delay Discounting Prizes	-.30	-.11	-.52	-.12
Weighted Average			**.25**	**.12**	**.33**	**.15**

Note. TB = Total Behavior scale. SB = Social Behavior subscale. AB = Academic Behavior subscale. EB = Emotional Behavior subscale. WIAT-III = Wechsler Individual Achievement Test, Third Edition. CBM = Curriculum-Based Measure. MAP = Missouri Assessment Program. CREVT = Comprehensive Receptive and Expressive Vocabulary Test. All statistics reported as Pearson correlation coefficients. Weighted averages calculated using absolute values.

[a]Standardized measures.

Table 7

Summary Findings from Studies Examining Relations to Other Variables (Social, Emotional, & Behavioral Measures)

Study	*n*	Measure	TB	SB	AB	EB
Eklund et al., 2017	1,044	Absences: Overall	-.19	-.17	-.18	-.14
Eklund et al., 2017	1,044	Absences: Within-Group	-.18	-.15	-.16	-.14
Eklund et al., 2017	1,044	Absences: Between-Group	-.20	-.24	-.23	-.08
von der Embse et al., 2018	1,158	Absences	-.19	-.12	-.20	-.16
Kilgus et al., 2016a (Study 1 [K-5])	864	BESS: Behavioral & Emotional Risk	-.93	-.79	-.86	-.72
Kilgus et al., 2016a (Study 1 [6-8])	864	BESS: Behavioral & Emotional Risk	-.94	-.85	-.88	-.69
Kilgus et al., 2016a (Study 2 [K-5])	1,534	BESS: Behavioral & Emotional Risk	-.94	-.85	-.88	-.75
Kilgus et al., 2016a (Study 2 [6-8])	1,534	BESS: Behavioral & Emotional Risk	-.94	-.83	-.88	-.72
Kilgus et al., 2018c	1,242	BESS: Behavioral & Emotional Risk	-.95	-.86	-.87	-.77
Kilgus et al., 2018c	1,242	BESS: Externalizing Problems	-.82	-.91	-.65	-.59
Kilgus et al., 2018c	1,242	BESS: School Problems	-.77	-.63	-.84	-.55
Kilgus et al., 2018c	1,242	BESS: Adaptive Skills	-.82	-.68	-.89	-.56
Kilgus et al., 2018c	1,242	BESS: Internalizing Problems	-.69	-.59	-.41	-.85
Stallone, 2021	106	CBCL Score (5-year-olds)	.00	.01	.07	-.11
Stallone, 2021	106	CBCL-TRF Score (6-year-olds)	-.04	.09	.01	-.04
Stallone, 2021	106	T-CRS: Task Orientation	.04	.05	-.04	.10
Stallone, 2021	106	T-CRS: Behavior Control	-.06	.01	-.13	-.01
Stallone, 2021	106	T-CRS: Assertiveness	.02	.02	-.05	.10
Stallone, 2021	106	T-CRS: Peer Social	.00	.03	-.10	.08
von der Embse et al., 2018	1,158	In-School Suspensions	-.05	-.05	-.03	-.02
von der Embse et al., 2018	1,158	Out-of-School Suspensions	-.28	-.34	-.17	-.23
Eklund et al., 2017	1,044	Office Discipline Referrals: Overall	-.33	-.39	-.25	-.25
Eklund et al., 2017	1,044	Office Discipline Referrals: Within-Group	-.33	-.38	-.24	-.25
Eklund et al., 2017	1,044	Office Discipline Referrals: Between-Group	-.36	-.42	-.28	-.27
von der Embse et al., 2018	1,158	Office Discipline Referrals: Minor Infractions	-.53	-.60	-.36	-.38
von der Embse et al., 2018	1,158	Office Discipline Referrals: Major Infractions	-.48	-.55	-.32	-.36
Whitley et al., 2019 (Time 1 EOY)	175	Office Discipline Referrals	-.39	-.54	-.16	-.28
Whitley et al., 2019 (Time 2 EOY)	175	Office Discipline Referrals	-.41	-.49	-.29	-.36
Tanner et al., 2018 (Time 1 [GenEd])	82	SDQ: Total Difficulties	-.95	-	-	-
Tanner et al., 2018 (Time 1 [SpEd])	82	SDQ: Total Difficulties	-.92	-	-	-
Tanner et al., 2018 (Time 2 [GenEd])	82	SDQ: Total Difficulties	-.93	-	-	-
Tanner et al., 2018 (Time 2 [SpEd])	82	SDQ: Total Difficulties	-.79	-	-	-
Roberson, 2019 (Time 1)	371	SDQ: Internalizing Behavior	-.63	-.39	-.45	-.79
Roberson, 2019 (Time 2)	332	SDQ: Internalizing Behavior	-.70	-.48	-.53	-.80
Roberson, 2019 (Time 1)	371	SDQ: Externalizing Behavior	-.84	-.88	-.73	-.52
Roberson, 2019 (Time 2)	332	SDQ: Externalizing Behavior	-.85	-.87	-.73	-.59
Roberson, 2019 (Time 1)	371	SDQ: Prosocial Behavior	.68	.66	.56	.51
Roberson, 2019 (Time 2)	332	SDQ: Prosocial Behavior	.70	.64	.61	.55
Kilgus et al., 2016b	346	SIBS: Internalizing Behavior	-.67	-.50	-.50	-.77
Kilgus et al., 2016b	346	SRSS: Internalizing & Externalizing Behavior	-.84	-.84	-.74	-.61
Roberson, 2019 (Time 1)	371	SWTRS: Psychopathology & Well-Being	.91	.73	.82	.75
Roberson, 2019 (Time 2)	332	SWTRS: Psychopathology & Well-Being	.91	.69	.87	.75
Weighted Average			**.58**	**.53**	**.50**	**.45**

Note. TB = Total Behavior scale. SB = Social Behavior subscale. AB = Academic Behavior subscale. EB = Emotional Behavior subscale. BESS-3 = Behavioral and Emotional Screening System. SRSS = Student Risk Screening Scale. SIBS = Student Internalizing Behavior Screener. SDQ = Strengths and Difficulties Questionnaire. SWTRS = Student Well-Being Teacher Report Scale. GenEd = General Education. SpEd = Special Education. T-CRS = Teacher-Child Rating Scale. CBCL = Child Behavior Checklist. CBCL-TRF = Child Behavior Checklist-Teacher Response Form. All statistics reported as Pearson correlations. Weighted averages calculated using absolute values.

Table 8

Summary Findings from Studies Examining Classification/Diagnostic Accuracy (Total Behavior Scale)

Study	*n*	Comparison Measure	Cut Score	AUC	SE	SP	PPV	NPV	PLR	NLR
Ezell, 2018	273	ADHD-RS	NR	.91	.92	.80	NR	NR	NR	NR
Hamsho, 2020	147	WIAT-III	NR	.65	.39	.77	.19	.90	NR	NR
Kilgus et al., 2016a (Study 1 [K-5])	864	BESS	36†	.97	.90	.93	.70	.98	12.86	.11
Kilgus et al., 2016a (Study 1 [6-8])	864	BESS	36†	.99	.95	.92	.66	.99	11.88	.05
Kilgus et al., 2016a (Study 2 [K-5])	1,534	BESS	36†	.98	.97	.93	.72	.95	13.86	.03
Kilgus et al., 2016a (Study 2 [6-8])	1,534	BESS	36†	.98	.88	.92	.63	.98	11.00	.13
Kilgus et al., 2016b	346	SRSS	47	.94	.88	.87	.72	.95	6.77	.14
Kilgus et al., 2016b	346	SIBS	42	.94	.90	.84	.36	.99	5.63	.12
Kilgus et al., 2018b (Time 1)	1243	BESS	27	.98	.91	.94	.55	.99	NR	NR
Kilgus et al., 2018b (Time 2)	704	BESS	27	.98	.86	.95	.58	.99	NR	NR
Kilgus et al., 2018c	1,242	BESS	36†	.98	.97	.88	.70	.99	9.08	.03
Kilgus et al., 2018d	704	BESS	NR	NR	.93	.91	.74	.98	10.33	.08
Roberson, 2019	371	SDQ	35	.95	.90	.86	.49	.98	NR	NR
Weighted Average				**.90**	**.91**	**.91**	**.62**	**.95**	**8.10**	**.05**

Note. †Cut scores denoted by instrument developers. NR = Not Reported. AUC = Area Under the Curve. SE = Sensitivity. SP = Specificity. PPV = Positive Predictive Value. NPV = Negative Predictive Value. PLR = Positive Likelihood Ratio. NLR = Negative Likelihood Ratio. BR = Base Rate. ADHD-RS = Attention Deficit Hyperactivity Disorder Rating Scale, Fourth Edition. WIAT-III = Wechsler Individual Achievement Test, Third Edition. BESS-3 = Behavioral and Emotional Screening System. SRSS = Student Risk Screening Scale. SIBS = Student Internalizing Behavior Screener. SDQ = Strengths and Difficulties Questionnaire.

Table 9

Summary Findings from Studies Examining Classification/Diagnostic Accuracy (Social Behavior Subscale)

Study	*n*	Comparison Measure	Cut Score	AUC	SE	SP	PPV	NPV	PLR	NLR
Kilgus et al., 2016a (Study 1 [K-5])	864	BESS	12†	.90	.81	.86	.51	.96	5.79	.22
Kilgus et al., 2016a (Study 1 [6-8])	864	BESS	12†	.96	.93	.85	.51	.99	6.20	.08
Kilgus et al., 2016a (Study 2 [K-5])	1,534	BESS	12†	.95	.86	.88	.56	.97	7.17	.16
Kilgus et al., 2016a (Study 2 [6-8])	1,534	BESS	12†	.92	.80	.90	.53	.97	7.90	.23
Kilgus et al., 2016b	346	SRSS	15	.93	.86	.87	.72	.94	6.62	.16
Kilgus et al., 2016b	346	SIBS	15	.85	.81	.71	.22	.97	2.79	.27
Kilgus et al., 2018b (Time 1)	1243	BESS	7	.96	.85	.93	.49	.99	NR	NR
Kilgus et al., 2018b (Time 2)	704	BESS	6	.97	.84	.93	.51	.99	NR	NR
Kilgus et al., 2018c	1,242	BESS	12†	.93	.88	.79	.55	.96	4.19	.15
Roberson, 2019	371	SDQ	11	.92	.86	.85	.47	.98	NR	NR
Weighted Average				**.93**	**.85**	**.86**	**.51**	**.97**	**4.64**	**.13**

Note. †Cut scores denoted by instrument developers. NR = Not Reported. AUC = Area Under the Curve. SE = Sensitivity. SP = Specificity. PPV = Positive Predictive Value. NPV = Negative Predictive Value. PLR = Positive Likelihood Ratio. NLR = Negative Likelihood Ratio. BR = Base Rate. BESS-3 = Behavioral and Emotional Screening System. SRSS = Student Risk Screening Scale. SIBS = Student Internalizing Behavior Screener. SDQ = Strengths and Difficulties Questionnaire.

Table 10

Summary Findings from Studies Examining Classification/Diagnostic Accuracy (Academic Behavior Subscale)

Study	*n*	Comparison Measure	Cut Score	AUC	SE	SP	PPV	NPV	PLR	NLR
Ezell, 2018	273	ADHD-RS	NR	.95	.84	.81	NR	NR	NR	NR
Kilgus et al., 2016a (Study 1 [K-5])	864	BESS	9†	.94	.91	.84	.52	.98	5.69	.11
Kilgus et al., 2016a (Study 1 [6-8])	864	BESS	9†	.95	.91	.83	.47	.98	5.35	.11
Kilgus et al., 2016a (Study 2 [K-5])	1,534	BESS	9†	.96	.93	.86	.55	.98	6.64	.08
Kilgus et al., 2016a (Study 2 [6-8])	1,534	BESS	9†	.94	.91	.85	.45	.99	6.07	.11
Kilgus et al., 2016b	346	SRSS	15	.88	.81	.79	.60	.92	3.86	.24
Kilgus et al., 2016b	346	SIBS	13	.85	.77	.73	.22	.92	2.85	.32
Kilgus et al., 2018b (Time 1)	1243	BESS	6	.93	.76	.90	.40	.98	NR	NR
Kilgus et al., 2018b (Time 2)	704	BESS	12	.94	.78	.91	.41	.98	NR	NR
Kilgus et al., 2018c	1,242	BESS	9†	.93	.83	.85	.62	.95	5.53	.20
Roberson, 2019	371	SDQ	9†	.87	.82	.80	.38	.97	NR	NR
Weighted Average				**.94**	**.86**	**.85**	**.47**	**.94**	**4.27**	**.10**

Note. †Cut scores denoted by instrument developers. NR = Not Reported. AUC = Area Under the Curve. SE = Sensitivity. SP = Specificity. PPV = Positive Predictive Value. NPV = Negative Predictive Value. PLR = Positive Likelihood Ratio. NLR = Negative Likelihood Ratio. BR = Base Rate. ADHD-RS = Attention Deficit Hyperactivity Disorder Rating Scale, Fourth Edition. SRSS = Student Risk Screening Scale. SIBS = Student Internalizing Behavior Screener. BESS-3 = Behavioral and Emotional Screening System. SDQ = Strengths and Difficulties Questionnaire.

Table 11

Summary Findings from Studies Examining Classification/Diagnostic Accuracy (Emotional Behavior Subscale)

Study	*n*	Comparison Measure	Cut Score	AUC	SE	SP	PPV	NPV	PLR	NLR
Kilgus et al., 2016a (Study 1 [K-5])	864	BESS	17†	.88	.90	.73	.38	.97	3.33	.14
Kilgus et al., 2016a (Study 1 [6-8])	864	BESS	17†	.86	.86	.73	.35	.97	3.19	.19
Kilgus et al., 2016a (Study 2 [K-5])	1,534	BESS	17†	.89	.88	.72	.36	.97	3.14	.17
Kilgus et al., 2016a (Study 2 [6-8])	1,534	BESS	17†	.88	.90	.65	.26	.98	2.57	.15
Kilgus et al., 2016b	346	SRSS	18	.84	.75	.77	.55	.89	3.26	.32
Kilgus et al., 2016b	346	SIBS	16	.97	.90	.86	.39	.99	6.43	.12
Kilgus et al., 2018b (Time 1)	1243	BESS	12	.90	.73	.90	.37	.98	NR	NR
Kilgus et al., 2018b (Time 2)	704	BESS	12	.89	.60	.91	.34	.97	NR	NR
Kilgus et al., 2018c	1,242	BESS	16	.91	.90	.73	.49	.96	3.33	.14
Roberson, 2019	371	SDQ	15	.87	.86	.73	.32	.97	NR	NR
Weighted Average				**.89**	**.84**	**.76**	**.37**	**.97**	**2.42**	**.12**

Note. †Cut scores denoted by instrument developers. NR = Not Reported. AUC = Area Under the Curve. SE = Sensitivity. SP = Specificity. PPV = Positive Predictive Value. NPV = Negative Predictive Value. PLR = Positive Likelihood Ratio. NLR = Negative Likelihood Ratio. BR = Base Rate. SRSS = Student Risk Screening Scale. SIBS = Student Internalizing Behavior Screener. BESS-3 = Behavioral and Emotional Screening System. SDQ = Strengths and Difficulties Questionnaire.

Table 12

Homogeneity Analysis of SAEBRS-TRS Coefficient Alpha & Correlation Coefficients

SAEBRS-TRS Scale/Subscale	Number of Coefficients	*k*	ES (CI)	*p*-value	*Q*	*p*-value	I^2
Coefficient Alpha							
Total Behavior	8	4	.94 (.93-.94)	< .0001	19.32	.0072	62.85%
Social Behavior	11	5	.93 (.90-.95)	< .0001	1859.89	< .0001	99.43%
Academic Behavior	12	6	.94 (.92-.96)	< .0001	2020.06	< .0001	99.52%
Emotional Behavior	11	5	.89 (.81-.93)	< .0001	3870.12	< .0001	99.71%
Correlation Coefficients (SEB)							
Total Behavior	27	6	.62 (.48-.73)	< .0001	1402.23	< .0001	98.67%
Social Behavior	19	6	.19 (.14-.24)	< .0001	34.78	< .0001	52.62%
Academic Behavior	19	6	.39 (.30-.48)	< .0001	136.40	< .0001	88.76%
Emotional Behavior	16	5	.28 (.25-.32)	< .0001	20.82	< .0001	15.38%
Correlation Coefficients (Academic)							
Total Behavior	31	6	.73 (.62-.81)	< .0001	3937.40	< .0001	99.45%
Social Behavior	34	5	.70 (.63-.76)	< .0001	2097.38	.1010	98.45%
Academic Behavior	34	5	.71 (.64-.77)	< .0001	2631.73	< .0001	98.78%
Emotional Behavior	37	6	.60 (.53-.67)	< .0001	1339.75	.1426	98.01%

Note. k = number of independent samples; ES = weighted coefficient alpha; CI = 95% confidence interval; *Q* = Cochran's heterogeneity *Q*-statistic with k-1 degrees of freedom; I^2 = heterogeneity index.

**p* < .05

Table 13

Moderator Analysis of SAEBRS-TRS Correlation Coefficients

Moderator	Number of Coefficients	*k*	ES (CI)	*p*-value	*Q*	*p*-value	I^2
Time of Administration							
Fall	34	5	.51 (.35-.64)	< .0001	12882.44	< .0001	99.73%
Winter	11	3	.20 (.09-.30)	.0004	70.20	< .0001	90.12%
Urbanicity							
Urban	44	6	.48 (.36-.58)	< .0001	6896.20	< .0001	99.33%
Suburban	7	2	.78 (.50-.91)	< .0001	402.90	< .0001	99.44%
Rural	15	4	.61 (.32-.79)	.0002	2657.11	< .0001	99.56%
Grade Level							
Elementary Only	62	12	.47 (.37-.56)	< .0001	10543.24	< .0001	99.31%
Secondary Only	15	4	.67 (.43-.82)	< .0001	4366.64	< .0001	99.73%

Note. k = number of independent samples; ES = weighted correlation coefficient; CI = 95% confidence interval;

Q = Cochran's heterogeneity *Q*-statistic with k-1 degrees of freedom; I^2 = heterogeneity index.

**p* < .05

Chapter 3

Study 2: Technical Adequacy of the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale: A Measurement Invariance Study

Abstract

School-based universal screening in the social, emotional, and behavioral (SEB) domains allows for the early identification of students in need of SEB support. Importantly, equitable assessment in universal screening for the SEB domains is critical to engage in accurate and ethical data-based decision-making. Measurement invariance is one method for examining potential inequities in assessment tools, permitting the ability to evaluate which assessments or assessment items perform differently across groups. As such, this study utilized multi-group confirmatory factor analysis to evaluate the extent of measurement invariance for a commonly used universal screening tool for the SEB domains: the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS). The sample consisted of 1,949 students in kindergarten through fourth grade in a Midwest, suburban school district. Examination of factor structures indicated the bifactor model yielded adequate fit and was utilized for measurement invariance testing. Multi-group confirmatory factor analysis results provided preliminary evidence that the SAEBRS-TRS displays invariance across a variety of student characteristics. Specifically, results supported configural and metric/scalar invariance of the bifactor model across the student characteristics of racial or ethnic identity, sex assigned at birth, and eligibility for free or reduced-price lunch. Yet, future research is needed to corroborate these findings. Limitations, implications for practice, and directions for future research are discussed.

Schools have become the primary provider of social, emotional, and behavioral (SEB) services for many children; yet estimates suggest that only one in eight schools conduct universal screening to identify students who may benefit from support (Bruhn et al., 2014). It is critical that schools engage in early identification of students in need of SEB support, especially considering the impact of the COVID-19 pandemic on student development across the SEB domains (Tang et al., 2021). Fortunately, universal screening may permit us to make accurate inferences about children's SEB functioning, provided they offer reliable and valid evidence. However, these tools may not function or perform equitably across students, indicating the potential bias of a measure. Specifically, bias in assessment refers to a systematic error in the measurement process that *differentially* influences scores for a particular group (Reynolds & Suzuki, 2013). As such, it is critical to evaluate for potential systematic differences to provide information regarding an assessment's applicability and accuracy for use with a variety of student populations (American Educational Research Association et al., 2014). Accordingly, the purpose of this study was to examine the potential assessment bias of the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS; Kilgus et al., 2016).

Equitable Assessment Practices

The *Standards for Educational and Psychological Testing* emphasize fairness and equity in testing as a critical component of the measurement development and validation process (American Educational Research Association et al., 2014). The *Standards* suggest that "those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population" (American Educational Research Association et al., 2014, p. 49). In accordance with these recommendations, instrument developers must continuously and comprehensively investigate various psychometric properties of the measure. One important component of this process is to determine the degree of equivalence or invariance across students (Rios & Wells, 2014). One method of examining the extent of a measure's invariance occurs through multi-group confirmatory factor analysis (MG-CFA; Dimitrov, 2010). The sequential process of MG-CFA includes the following steps: (a) establishing a baseline model; (b) evaluating model fit of the baseline model; and (c) evaluating models of increasingly restrictive equality constraints in comparison to the baseline model (Rios & Wells, 2014). If the model fit changes minimally as equality constraints are included, the measure is considered equivalent or invariant across groups (Brown, 2015). This means that both constructs and measures operate similarly across various sample characteristics, such as racial or ethnic identity or sex assigned at birth (Vandenberg & Lance, 2000).

Traditional invariance testing through MG-CFA consists of three main levels of invariance: configural, metric, and scalar. Configural invariance is defined as the degree to which the same items are measuring the same factors across groups, indicating the same items load onto the same factors in both groups (Byrne, 2012; Pendergast et al., 2017). Metric invariance, also known as weak invariance, refers to the degree of the relationships between items and factors across groups (Byrne, 2012; Pendergast et al., 2017). If the relationships between items and factors are significantly different across groups, the measure is considered non-invariant (Byrne, 2012; Pendergast et al., 2017). Scalar invariance, also known as strong invariance, refers to cases of both groups with similar scores on the latent factor having an equal probability of shifting between response categories (Byrne, 2012; Pendergast et al., 2017). However, recent findings indicate MG-CFA with categorical/ordinal data cannot be conducted in the order it is conducted with continuous/interval data (Wu & Estabrook, 2016). Alternatively, after configural invariance is tested, it is recommended to constrain the item factor loadings and item thresholds to be equivalent across groups (Muthén & Muthén, 2020), analyzing the metric and scalar invariance jointly. Taken together, evaluation of the configural and metric/scalar invariance indicates the degree to which the measure is invariant across groups, a critical and often overlooked component of the validation process.

The Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS)

Many brief behavior rating scales have been developed to detect and identify students who may benefit from SEB support. One of these commonly used scales is the Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS; Kilgus et al., 2016). The SAEBRS-TRS is a commercially published measure available via the FastBridge suite of assessments. This rating scale assesses student needs across the domains of social, academic, and emotional behavior. The SAEBRS-TRS factor structure has been examined and supported through exploratory factor analysis (EFA) and confirmatory factor analysis (CFA; Kilgus et al., 2013; Kilgus et al., 2018a; von der Embse et al., 2016; von der Embse et al., 2019). Notably, EFA was only conducted with the original version of the measure (i.e., SABRS), which provided support for a bifactor structure inclusive of one general factor (Total Behavior) and two specific factors (Social Behavior and Academic Behavior; Kilgus et al., 2013). Two studies conducted CFA with the original SABRS (Kilgus et al., 2015; Pendergast et al., 2017). Additionally, two studies conducted CFA with the revised measure (SAEBRS; Kilgus et al., 2018a; von der Embse et al., 2016), with both confirming support for a bifactor model. Further, studies have supported the internal consistency, test-retest reliability, convergent validity, discriminant validity, concurrent validity, and classification/diagnostic accuracy of scores (see Study 1). Despite these strengths, a single research group conducted the majority of studies (i.e., Kilgus, von der Embse, and colleagues). While the SAEBRS-TRS has displayed adequate psychometric properties within these studies, the extent to which the measure is appropriate for use with students across various characteristics (e.g., racial or ethnic identity, sex assigned at birth, eligibility for free or reduced-price lunch) has yet to be systematically evaluated.

Indeed, a single measurement invariance study was conducted for the original version of the measure (i.e., SABRS), examining invariance across White and Black students (Pendergast et al., 2017). However, this version of the measure is fundamentally different from the revised version and results cannot be considered indicative of the internal structure of the SAEBRS-TRS. Additionally, one measurement invariance study was conducted using the revised version (SAEBRS-TRS), with findings providing preliminary support for invariance across age, sex assigned at birth, and eligibility for special education services (von der Embse et al., 2019). Yet, the study combined the student- and teacher-report versions for analysis (i.e., mySAEBRS and SAEBRS-TRS), leaving scrutiny of the teacher-report version inexecutable. Given widespread use of the SAEBRS-TRS specifically, understanding this information is critical.

Purpose

The increased need for early prevention and intervention services has resulted in nation-wide recommendations for educators to prioritize identification of *all* students who might be considered at-risk for SEB problems. As schools and districts consider the adoption of various universal screening tools, attention to the appropriateness and usability for the intended population is warranted. As such, the primary aim of this study is to examine the degree of measurement invariance of the SAEBRS-TRS. While extant validity evidence is promising, the potential invariance of the measure has yet to be examined exclusively for the revised version. This is problematic as the assessment is widely used in schools across the United States, despite the lack of evidence of its interpretability across said characteristics. This study aims to examine the invariance of the SAEBRS-TRS across student groups based on racial or ethnic identity, sex assigned at birth, and eligibility for free or reduced-price lunch (FRL). Primary research questions are as follows:

1. To what extent do SAEBRS-TRS scores exhibit internal consistency within the study sample?
2. To what extent is concurrent validity evidence demonstrated when comparing SAEBRS-TRS scores and concurrent measures of student functioning (i.e., Office Discipline Referrals, attendance)?
3. To what extent does the SAEBRS-TRS exhibit measurement invariance when used with a variety of student characteristics (i.e., racial or ethnic identity, sex assigned at birth, eligibility for free or reduced-price lunch)?

Method

This study was registered through Open Science (doi: 10.17605/OSF.IO/UV9CP) prior to data collection in fall of 2021.

Participants

Data collection was conducted in a suburban, Midwestern school district in October of 2021. The sample included 169 teachers of 1,949 students across four elementary schools (kindergarten through fourth grade). The majority of students were White (*n* = 972; 49.9%) and male (*n* = 993; 50.9%). A moderate proportion of students identified as eligible for free or reduced-price lunch (*n* = 852; 43.7%), Multilingual Learner services (*n* = 353; 18.1%), or special education services (*n* = 234; 12.0%). Student characteristics are presented in Table 1. The majority of teachers who completed the SAEBRS-TRS were White (*n* = 135; 79.9%) and female (*n* = 153; 90.5%). Teacher characteristics are presented in Table 2.

Measures

The Social, Academic, and Emotional Behavior Risk Screener - Teacher Rating Scale (SAEBRS-TRS)

The SAEBRS-TRS is a brief universal screening tool consisting of 19 items, which aims to identify students at-risk for SEB problems. The SAEBRS-TRS includes one full scale (Total Behavior [TB]) consisting of three subscales (Social Behavior [SB], Academic Behavior [AB], Emotional Behavior [EB]). The SAEBRS-TRS is completed via paper-and-pencil or online administration. In addition to teacher-report, the measure has parent- and student-report versions (SAEBRS-PRS and mySAEBRS, respectively). To complete the SAEBRS-TRS, teachers rate the frequency a student engaged in various adaptive and maladaptive behaviors in the previous month. Ratings are completed using a four-point Likert scale (0 = *never*, 1 = *sometimes*, 2 = *often*, 3 = *almost always*). The six Social Behavior subscale items (e.g., Temper Outbursts) are scored from zero to 18; the six Academic Behavior items (e.g., Academic Engagement) are scored from zero to 18; and the seven Emotional Behavior items (e.g., Fearfulness) are scored from zero to 21. Negatively worded items (e.g., Disruptive Behavior) are reverse-scored. A total score is calculated by adding the subscale scores together, ranging from zero to 57. Subsequently, higher scores indicate more adaptive student behavior and lower scores indicate more maladaptive student behavior.

Student & Teacher Variables

Information regarding both student and teacher characteristics and variables were collected. Student information included the following: grade, sex assigned at birth, racial or ethnic identity, eligibility for Multilingual Learner services, eligibility for free or reduced-price lunch, eligibility for special education services, Office Discipline Referrals (ODRs), and attendance data. Teacher information included the following teacher-level characteristics: gender identity, racial or ethnic identity, average number of years teaching, and highest earned degree.

Office Discipline Referrals. Office Discipline Referrals (ODRs) were used as a criterion-related measure of student functioning to examine the relation of SAEBRS-TRS scores to other indicators of student behavior. ODRs represent the total number of times a teacher called for behavior support for a student, as reported by the school district as minor infractions (e.g., work refusal) or major infractions (e.g., disorderly conduct). ODR data was collected for the entirety of the first trimester of the 2021-2022 school year.

Attendance. Attendance was used as a criterion-related measure of student functioning to examine the relation of SAEBRS-TRS scores to other indicators of student behavior. Attendance represented the total number of days absent during the current school year. Attendance data was collected for the entirety of the first trimester of the 2021-2022 school year.

Procedures

Deidentified data were used for analysis; thus, this study was determined by the Institutional Review Board to not constitute human research. Administration of the SAEBRS-TRS occurred in October of 2021, consistent with instrument developer recommendations to implement the screener six weeks into the school year. All teachers completed the SAEBRS-TRS electronically through the FastBridge suite of assessments. The deidentified dataset was provided by the district in December of 2021, inclusive of data collected from the aforementioned measures.

Data Analysis

Two statistical packages were used for this study. The Statistical Package for the Social Sciences (SPSS) Version 27.0 was used to calculate descriptive statistics and correlational analyses. M*plus* Version 8.5 was used to analyze confirmatory factor analyses (Muthén & Muthén, 2020).

Preliminary Analyses

First, preliminary data screening was conducted to evaluate statistical assumptions, including the identification of outliers, distribution of variables, and missing data. Item-level analyses were examined, including response distributions and point-biserial correlations.

Estimation Method

While maximum likelihood estimation is a commonly used estimation technique for MG-CFA, it is not recommended for use with non-normal and ordinal data (Bollen, 1989). Therefore, in accordance with best practice and prior recommendations, weighted least squares with mean and variance adjustment (WLSMV) estimation was employed to address the ordinal nature of data (Beauducel & Herzberg, 2006). In factor analytic techniques, the relationships between continuous variables are obtained from product-moment correlations or covariances, which do not accurately reflect the variance in ordinal data (Pendergast et al., 2017). Therefore, estimators such as WLSMV are recommended for factor analysis with ordinal data as they are based on polychoric matrices, consequently yielding more accurate results compared to those based on raw scores (Flora et al., 2012; Forero et al., 2009; Pendergast et al., 2017).

Missing Data

For the present study, there was no missing data in the SAEBRS-TRS dataset. Within the online platform, respondents must complete all items to submit the SAEBRS-TRS; missing responses are not allowed. Therefore, no missing data techniques were employed.

Primary Analyses

Internal Consistency. The internal consistency of scores was calculated for the full scale and subscales. Although coefficient alpha is the most commonly used measure of internal consistency, it is generally a poor indicator due to its insensitivity to item multidimensionality (Sijtsma, 2009). Therefore, hierarchical omega coefficients (ω_H) were calculated to examine the extent to which total scores are indicative of a common variable. When interpreting hierarchical omega, it is suggested larger values represent a single variable (i.e., unidimensionality). In addition to hierarchical omega, subscale omega (ω_s) was examined for specific factors to identify variance beyond the general factor. In accordance with previously specified interpretive guidelines, hierarchical omega and subscale omega values greater than .70 are considered acceptable (Reise et al., 2012; Rios & Wells, 2014). Results were also interpreted in relation to previous findings for the internal consistency of the SAEBRS-TRS.

Concurrent Validity Evidence. Measures of concurrent validity evidence were calculated for all subscales and the full scale. Concurrent validity evidence was examined to explore the relationship between SAEBRS-TRS scores and extant outcome data (i.e., ODRs, attendance). This included the calculation of Spearman correlation coefficients, due to the ordinal nature of the variables. Correlation coefficients between .01 and .19 were considered negligible, .20 to .29 were considered weak, between .30 and .39 were considered moderate, between .40 and .69 were considered strong, and greater than .70 were considered very strong (Dancey & Reidy, 2007). Additionally, coefficients were interpreted in relation to past SAEBRS-

TRS research findings and similar measures of SEB domains. Specifically, findings were interpreted using established criteria and in comparison to the Behavioral and Emotional Screening System, Third Edition (BESS-3; Kamphaus & Reynolds, 2015). The BESS-3 is a similar measure to the SAEBRS-TRS, as it is a common universal screening tool used in schools for the SEB domains.

Measurement Invariance. Multi-group confirmatory factor analysis (MG-CFA) was used to evaluate measurement invariance. First, a series of CFAs was conducted inclusive of the full sample to examine various model structures and establish a baseline model. The following criteria, established *a priori*, were considered indicative of acceptable model fit: Root Mean Square Error of Approximation (RMSEA) ≤ .08; Confirmatory Fit Index (CFI) ≥ .90; Tucker-Lewis fit index (TLI) ≥ .90; Standardized Root Mean Squared Residual (SRMR) ≤ .08; and standardized factor loadings ≥ .30 (Brown, 2015; Browne & Cudeck, 1993; Hu & Bentler, 1999; Kline, 2010). Next, separate CFAs were conducted for the following categories: racial or ethnic identity (White, Hispanic or Latino, Asian, Black or African American, and Two or More Races), sex assigned at birth (female, male), and eligibility for free or reduced-price lunch (receiving free or reduced-price lunch, not receiving free or reduced-price lunch). It should be noted these categories were used as reported by the district, which include the federal definitions of racial or ethnic identity. For example, the present study used the term “Hispanic or Latino” as a category of racial or ethnic identity. It should be noted that a number of alternatives are preferred (e.g., Latinx, Latine, Latino/a) to represent those originating from Latin America. This study’s categorization of racial and ethnic identity has inherent limitations, which are further addressed in the discussion section of this manuscript.

The invariance of the bifactor structure was tested by placing increasingly restrictive equality constraints to evaluate each model's configural invariance and metric/scalar invariance. Configural invariance testing consisted of conducting CFA across groups simultaneously, with all parameters freely estimated. Next, in accordance with best practices (Wu & Estabrook, 2016; Muthén & Muthén, 2020), metric/scalar invariance was tested by building onto the previous model and restricting both the factor loadings and intercepts to be equal across groups. Each nested model was compared to the more restrictive model using the change in x^2 (Δx^2), change in CFI (ΔCFI), and change in RMSEA (ΔRMSEA) values (Pendergast et al., 2017). The following guidelines were established *a priori* to evaluate the degree to which the more restrictive model had a comparable fit to the less restrictive model: non-significant Δx^2 ($p > .05$), ΔCFI < .01, and ΔRMSEA < .015 (Byrne, 2012; Meade et al., 2008; Pendergast et al., 2017; Satorra & Bentler, 2001). This process was conducted for each group included in the analysis. Similar to the interpretive guidelines for concurrent validity evidence, findings were interpreted based on the aforementioned criteria and in comparison to the Behavioral and Emotional Screening System, Third Edition (BESS-3; Kamphaus & Reynolds, 2015).

Results

Descriptive Statistics

Preliminary data screening demonstrated that means and standard deviations for all items were within the expected range. Specifically, means ranged from 2.02 to 2.84 and standard deviations ranged from .42 to .95. However, all items had non-normal distributions with negative skewness and negative kurtosis. Item skewness ranged from -0.10 to -2.72 and item kurtosis ranged from -17.87 to -69.70. To account for the non-normal distribution as well as the ordinal nature of variables, WLSMV estimation was employed (Brown, 2015). There was no missing data identified in the dataset; therefore, procedures for missing data handling were unnecessary. Descriptive statistics are provided in Table 3.

Internal Consistency

Hierarchical omega was used to calculate the internal consistency of the general factor and specific factors. According to the predetermined threshold of .70 indicating acceptability (Reise et al., 2012; Rios & Wells, 2014), hierarchical omega indicated acceptable internal consistency for the TB scale (.937). Additionally, subscale omega indicated acceptable internal consistency for the SB subscale (.890), AB subscale (.921), and EB subscale (.846).

Concurrent Validity Evidence

Spearman correlation coefficients were used to examine concurrent validity estimates between the SAEBRS-TRS and attendance and Office Discipline Referrals. For attendance, all coefficients were statistically significant across scales (p < .05). Correlation coefficients were -.123 for the TB scale, -.060 for the SB subscale, -.159 for the AB subscale, and -.065 for the EB subscale. For ODRs classified as minor infractions, all coefficients were statistically significant across scales (p < .05). Correlation coefficients were -.164 for the TB scale, -.186 for the SB subscale, -.137 for the AB subscale, and -.124 for the EB subscale. For ODRs classified as major infractions, all coefficients were statistically significant across scales (p < .05). Correlation coefficients were -.253 for the TB scale, -.278 for the SB subscale, -.211 for the AB subscale, and -.204 for the EB subscale. Table 4 contains Spearman correlation coefficients for all scales.

Multi-Group Confirmatory Factor Analysis

Baseline Confirmatory Factor Analysis

Prior to examining measurement invariance, CFAs were conducted to determine the most appropriate baseline model. Analyses were conducted for a unidimensional model, second-order model, and bifactor model. The following fit statistics were observed for the unidimensional model: x^2 (152) = 7456.33, $p < .05$; RMSEA = .157; CFI = .883; TLI = .869; SRMR = .131. Standardized factor loadings for the unidimensional model ranged from .560 to .901 for the general factor. Next, the following fit statistics were observed for the second-order model: x^2 (149) = 4518.60, $p < .05$; RMSEA = .123; CFI = .930; TLI = .920; SRMR = .089. Standardized factor loadings for the second-order model ranged from .789 to .930 for the TB subscale, .832 to .894 for the SB subscale, .799 to .931 for the AB subscale, and .647 to .927 for the EB subscale. Last, the following fit statistics were observed for the bifactor model: x^2 (133) = 2227.52, $p < .05$; RMSEA = .090; CFI = .967; TLI = .957; SRMR = .053. Standardized factor loadings for the bifactor model ranged from .353 to .895 for the TB scale, -.138 to .526 for the SB subscale, .310 to .662 for the AB subscale, and .234 to .843 for the EB subscale. Two factor loadings on the SB subscale were negative (Item 2 and Item 5). Despite the negative factor loadings on the SB subscale of the bifactor model, it was determined to move forward with this model due to the optimal model fit statistics and consistency with prior research for the SAEBRS-TRS. Table 5 provides an overview of model fit statistics and factor loadings.

Single-Group Confirmatory Factor Analysis

Single-group CFAs were conducted separately for each group to ensure adequate model fit for each subsample. A total of nine single-group CFAs were conducted: two based on sex assigned at birth (female, male), five based on racial or ethnic identity (White, Black, Asian, Two or More Races, Hispanic or Latino), and two based on eligibility for free or reduced-price lunch (FRL, Non-FRL). Table 6 provides an overview of model fit statistics for each subsample.

Sex Assigned at Birth. Fit statistics for the female CFA were as follows: x^2 (133) = 1078.59, $p < .05$; RMSEA = .086; CFI = .964; TLI = .954; SRMR = .063. Fit statistics for the male CFA were as follows: x^2 (133) = 1318.87, $p < .05$; RMSEA = .095; CFI = .968; TLI = .959; SRMR = .052. Overall, both models displayed adequate fit and were similar across groups. Therefore, it was deemed appropriate to move forward with invariance testing.

Racial or Ethnic Identity. Fit statistics for each CFA were as follows: White (x^2 [133] = 1053.53, $p < .05$; RMSEA = .084; CFI = .968; TLI = .959; SRMR = .052), Black (x^2 [133] = 392.73, $p < .05$; RMSEA = .092; CFI = .976; TLI = .970; SRMR = .058), Asian (x^2 [133] = 290.51, $p < .05$; RMSEA = .081; CFI = .973; TLI = .966; SRMR = .090), Two or More Races (x^2 [133] = 291.88, $p < .05$; RMSEA = .084; CFI = .977; TLI = .971; SRMR = .058), and Hispanic or Latino (x^2 [133] = 593.54, $p < .05$; RMSEA = .099; CFI = .963; TLI = .952; SRMR = .080). Overall, all five models displayed adequate fit and were similar across groups. Therefore, it was deemed appropriate to move forward with invariance testing.

Eligibility for Free or Reduced-Price Lunch. Fit statistics for the CFA of students receiving FRL were as follows: x^2 (133) = 1135.13, $p < .05$; RMSEA = .094; CFI = .967; TLI = .957; SRMR = .057. Fit statistics for the CFA of students not receiving FRL were as follows: x^2 (133) = 1216.19, $p < .05$; RMSEA = .086; CFI = .965; TLI = .956; SRMR = .053. Overall, both models displayed adequate fit and were similar across groups. Therefore, it was deemed appropriate to move forward with invariance testing.

Measurement Invariance

Multi-group confirmatory factor analyses (MG-CFA) were used to examine measurement invariance. As recommended by prior research, if there were more than two groups, the referent group for each model was the group with the largest sample size (Fischer & Karl, 2019). Table 6 provides an overview of invariance testing results.

Sex Assigned at Birth. Configural invariance testing resulted in the following fit statistics: x^2 (300) = 2404.49, $p < .05$; RMSEA = .085; CFI = .967; TLI = .962; SRMR = .058. The metric/scalar model was then tested, resulting in the following fit statistics: x^2 (323) = 2862.03, $p < .05$; RMSEA = .090; CFI = .960; TLI = .958; SRMR = .062. The change in RMSEA (△RMSEA = .005) and change in CFI (△CFI = .007) reflected that constraining the model did not significantly adjust the model fit and invariance was supported. However, the change in chi-square (△x^2 [*df*] = 441.97 [23], $p < .05$) was statistically significant, indicating potential non-invariance.

Eligibility for Free or Reduced-Price Lunch. Configural invariance testing resulted in the following fit statistics: x^2 (300) = 2467.96, $p < .05$; RMSEA = .085; CFI = .965; TLI = .961; SRMR = .058. The metric/scalar model was then tested, resulting in the following fit statistics: x^2 (323) = 2384.21, $p < .05$; RMSEA = .081; CFI = .967; TLI = .965; SRMR = .058. The change in RMSEA (△RMSEA = .004) and change in CFI (△CFI = .002) reflected that constraining the model did not significantly adjust the model fit and invariance was supported. However, the change in chi-square (△x^2 [*df*] = 93.64 [23], $p < .05$) was statistically significant, indicating potential non-invariance.

Racial or Ethnic Identity. Configural invariance testing resulted in the following fit statistics: x^2 (801) = 2566.53, $p < .05$; RMSEA = .076; CFI = .973; TLI = .971; SRMR = .064. The metric/scalar model was then tested, resulting in the following fit statistics: x^2 (893) = 2786.40, $p < .05$; RMSEA = .074; CFI = .971; TLI = .972; SRMR = .067. The change in RMSEA (△RMSEA = .002) and change in CFI (△CFI = .002) reflected that constraining the model did not significantly adjust the model fit and invariance was supported. However, the change in chi-square (Δx^2 [*df*] = 438.38 [92], $p < .05$) was statistically significant, indicating potential non-invariance.

Discussion

Given the increase in universal screening in the SEB domains, sufficient validity evidence is critical to engage in equitable screening practices. Of particular importance, measurement invariance demonstrates the extent to which items are measuring the intended construct among the intended population (American Educational Research Association et al., 2014). As such, before inferences can be made on the basis of SAEBRS-TRS scores, potential invariance must be thoroughly examined. Moreover, measurement invariance is a necessary component in the evaluation of potential bias, considering the racial, ethnic, and gender disproportionality in the identification of students in need of SEB supports and services (Kalberg et al., 2011; Raines et al., 2012; Skiba et al., 2002). In consideration of the need for universal screening measures with minimal bias, the purposes of this study were three-fold: (a) to examine the extent SAEBRS-TRS scores exhibit internal consistency within the study sample; (b) to examine the relationship between SAEBRS-TRS scores and concurrent measures of student functioning (i.e., Office Discipline Referrals, attendance); and (c) to examine the extent the SAEBRS-TRS exhibits measurement invariance for students with various characteristics (i.e., racial or ethnic identity, sex assigned at birth, eligibility for free or reduced-price lunch).

Research Question #1

First, we examined the internal consistency of SAEBRS-TRS scores. The TB scale displayed acceptable internal consistency based on guidelines in extant literature ($\omega_H \geq .70$; Reise et al., 2012) and previous research conducted for the SAEBRS-TRS (Kilgus et al., 2018b). Considering the value of hierarchical omega can be interpreted as the amount of variance attributable to the general factor (McDonald, 1999), this indicates that score variance is largely attributable to the TB scale. As such, it is recommended that interpretation of the SAEBRS-TRS be conducted using the TB scale.

Research Question #2

Second, we examined the relationship between SAEBRS-TRS scores, attendance, and Office Discipline Referrals (ODRs). Despite Spearman correlations being statistically significant across all scales, the strength of relationships were negligible to weak for both attendance and ODRs with SAEBRS-TRS scales. Specifically, all correlations between the SAEBRS-TRS scales and attendance were negligible, all correlations between the SAEBRS-TRS scales and ODRs classified as minor infractions were negligible, and all correlations between the SAEBRS-TRS scales and ODRs classified as major infractions were weak. These findings are lower than expected when compared to prior SAEBRS-TRS research, which displayed weak to moderate relationships with ODRs, with correlations ranging from -.16 to -.54 (Eklund et al., 2016; Kilgus et al., 2017; Whitley & Cuenca-Carlino, 2019). No prior studies using the SAEBRS-TRS have examined the association with attendance.

Additionally, weaker relationships with attendance and ODRs were identified in this study in comparison to findings for the BESS-3 (Naser et al., 2018). Concurrent correlations between the BESS-3 and attendance were moderate (Pearson r = .337), and correlations between the BESS-3 and ODRs were small (Pearson r = .272). This discrepancy in SAEBRS-TRS and BESS-3 findings is important, especially considering the frequency schools use ODRs to measure student need for SEB support. Although ODRs may provide information on the frequency a student engages in behavior not accepted by school expectations or guidelines, it is not necessarily a strong indicator of SEB risk. Universal screening tools, in conjunction with other data points, are promising for providing more objective and accurate information related to student SEB needs.

Research Question #3

Third, we employed multi-group confirmatory factor analysis (MG-CFA) to examine the extent of measurement invariance of the SAEBRS-TRS across various student characteristics. Importantly, the single-group CFA results of this study suggest the factor loadings and fit indices are similar to those reported in prior research (Kilgus et al., 2018a; von der Embse et al., 2016). This is true for both the baseline model inclusive of all participants, as well as the single-group CFAs for separate student characteristics. It is important to note that two items (Item 2 and Item 5) displayed negative loadings to their respective subscale (SB) within the bifactor model. This indicates the two items are stronger measures of the general factor (TB) in comparison to the specific factor (SB). This finding is not entirely unexpected, as items within a bifactor model often demonstrate a stronger association with the general factor compared to specific factors. Specifically, non-significant and small factor loadings on a specific factor are an indicator the relationship between the item and specific factor is negligible, and interpretation of the item should be in regard to the general factor (Reise et al., 2012; Rios & Wells, 2014). This finding corroborates prior research of the SAEBRS-TRS, as Kilgus et al. (2018a) findings indicated weak factor loadings related to the specific factor of SB, specifically for Item 2 (.01) and Item 5 (-.01). Additionally, von der Embse et al. (2016) employed a bifactor model, with standardized factor loadings the weakest for Item 2 (.41) and Item 5 (.45), although they were much stronger than findings for the present study.

Results of the MG-CFA provided support for the use of the SAEBRS-TRS across various student characteristics. Configural and metric/scalar invariance were established across racial or ethnic identity, sex assigned at birth, and eligibility for free or reduced-price lunch, indicating the measure functions similarly across said characteristics. Specifically, two change-in-model-fit indices (change in CFI and change in RMSEA) indicated invariance across all three models. However, changes in chi-square estimates were statistically significant across all three models, indicating potential non-invariance. This should be interpreted cautiously as chi-square is sensitive to sample size, resulting in a high likelihood of statistical significance in well-fitting models due to the large sample size required for CFA techniques (Kline, 2010). Additionally, prior research suggests that in MG-CFA with sample sizes larger than 200, significant change in chi-square values are likely negligible (Meade et al., 2008).

Limitations & Future Directions

Despite the promising findings of the present study, there are several limitations to note. First, it is generally recommended that MG-CFA be conducted with a minimum of 200 participants per group (Gonzalez-Roma et al., 2006). Although small sample sizes were identified for some racial or ethnic identities (i.e., Asian; Two or More Races), the authors decided to include them in the MG-CFA despite being smaller than the recommendations of minimum 200 participants per group. This decision was made due to sample sizes being very close to the recommended minimum and in an effort to more appropriately represent the U.S. student population. Because smaller samples have less power to detect statistical differences, a null result would be more likely in this case. Conversely, the results of the analyses in this study suggested statistical differences, despite the sample size. Future research should aim to include sufficient sample sizes to substantiate the findings of the present study.

Additionally, the current study was conducted with a limited age group (kindergarten through fourth grade) in a suburban, Midwestern school district. While findings are informative, they may not be reflective of populations with different student characteristics. This is particularly important as most SAEBRS-TRS research has been conducted in the Midwest and may not adequately reflect SEB constructs across the nation. Moreover, teacher characteristics were not representative of student characteristics, with a relatively homogeneous teacher population across racial or ethnic identity and gender identity, and a relatively diverse student population across racial or ethnic identity and sex assigned at birth. It is important for future research to replicate these findings, and to examine the extent these findings hold across different settings (e.g., rural and urban) with a variety of age levels (e.g., secondary).

Inherent limitations exist in the categorization of racial and ethnic identities utilized in this study. This analysis does not take into account within-group differences, as response options were not designed to gain that level of specificity. Ideally, response options would have allowed a greater degree of specificity so as to better represent identities. For example, allowing for regional identification of individuals of Asian ancestry or Asian descent (e.g., East Asian, South Asian). This is a critical adjustment needed across racial or ethnic identities, as people of Asian, African, and Indigenous descent have been restricted by the identification guidelines of the dominant culture. Moreover, there are clear issues with the identification of individuals of Two or More Races into a single category, resulting in the homogenization of different cultures and experiences. Because of this, it is not possible to examine the impact of the variety of cultures and lived experiences captured within this single category. Future research must be effortful in disaggregating categories of racial or ethnic identity for appropriate representation of those involved in the research.

Further, race and ethnicity are not synonymous. Race refers to “physical differences that groups and cultures consider socially significant”, while ethnicity refers to “shared cultural characteristics such as language, ancestry, practices, and beliefs" (American Psychological Association, 2020). Yet, the categorizations used do not allow for individuals to identify themselves fully, as race and ethnicity were grouped into a single category. For example, one could not identify as both Black and Hispanic or Latino, as only one response option was able to be selected. Future analyses should address this limitation and gather more accurate and representative categorizations of racial or ethnic identity.

Lastly, interpretive issues related to the model specification of the SAEBRS-TRS remain prevalent. The underlying theory of the SAEBRS-TRS is one of interrelated social, emotional, and behavioral domains. Yet, the bifactor model assumes that the general and specific factors (i.e., full scale and subscales) are orthogonal or uncorrelated with each other. This is problematic, as the model specification does not match the theoretical basis of the measure. The present study chose to examine the bifactor model based on prior research and the comparison of fit statistics for various models (e.g., unidimensional model, second-order model). However, it is suggested that future research continue to investigate the second-order model. At a minimum, future studies should compare the fit of the second-order model and bifactor model prior to employing measurement invariance techniques.

Conclusion

Measurement invariance is a frequently overlooked component of measure development, as evidenced by the scarcity of invariance studies of widely used universal screening tools for the social, emotional, and behavioral domains. Considering the continued injustices of marginalized students, disproportionate identification of students eligible for special education services, and increase in the need for SEB services and supports following the COVID-19 pandemic, it is critical for educators to utilize equitable tools to identify students warranting support. This is the first study to examine measurement invariance of the revised SAEBRS-TRS measure, providing preliminary evidence of the measure's equivalence across the student characteristics of racial or ethnic identity, sex assigned at birth, and free or reduced-price lunch. Although the current study displays promising evidence of invariance across said groups, additional evidence is necessary to corroborate the findings and firmly establish invariance.

Table 14

District-Reported Characteristics of Student Participants

Characteristic	*N*	%
Sample Size	1949	100
Grade		
Kindergarten	344	17.7
First Grade	441	22.6
Second Grade	407	20.9
Third Grade	402	20.6
Fourth Grade	355	18.2
Sex Assigned at Birth		
Female	956	49.1
Male	993	50.9
Racial or Ethnic Identity		
White	972	49.9
Black or African American	233	12.0
Asian	181	9.3
Two or More Races	170	8.7
Hispanic or Latino	355	18.2
Other	38	1.9
Eligibility for Special Education		
Receiving Special Education Services	234	12.0
Not Receiving Special Education Services	1715	88.0
Eligibility for Free or Reduced-Price Lunch		
Receiving Free or Reduced-Price Lunch	852	43.7
Not Receiving Free or Reduced-Price Lunch	1097	56.3
Multilingual Learner		
Receiving Multilingual Learner Services	353	18.1
Not Receiving Multilingual Learner Services	1596	81.9

Table 15

District-Reported Characteristics of Teacher Participants

Characteristic	*N*	%
Sample Size	169	100
Sex Assigned at Birth		
Female	153	90.5
Male	16	9.5
Racial or Ethnic Identity		
White	135	79.9
Hispanic or Latino	18	10.6
Black or African American	5	3.0
Asian	11	6.5
Two or More Races	0	0
Other	0	0
Highest Level of Education		
Master’s Degree	59	34.9
Bachelor’s Degree	38	22.5
Not Reported	72	42.6
Years of Teaching Experience		
1-3 Years	41	24.3
4-6 Years	20	11.8
7-9 Years	24	14.2
10+ Years	12	7.1
Not Reported	72	42.6

Table 16

Item-Level Descriptive Statistics for SAEBRS-TRS Scores

SAEBRS-TRS Item/Scale	Mean	SD	Skewness	Kurtosis
Item 1	2.61	.68	-1.78	-37.50
Item 2	2.42	.76	-0.10	-20.83
Item 3	2.76	.60	-2.72	-65.74
Item 4	2.44	.79	-1.37	-29.29
Item 5	2.50	.73	-1.25	-24.43
Item 6	2.38	.85	-1.31	-27.07
Item 7	2.36	.78	-0.89	-19.94
Item 8	2.34	.84	-1.02	-21.54
Item 9	2.30	.83	-0.84	-19.07
Item 10	2.21	.95	-1.02	-21.53
Item 11	2.02	.88	-0.70	-20.45
Item 12	2.31	.79	-0.75	-17.86
Item 13	2.68	.57	-1.78	-39.73
Item 14	2.84	.42	-2.72	-69.70
Item 15	2.40	.83	-1.16	-23.42
Item 16	2.52	.67	-1.26	-25.79
Item 17	2.64	.59	-1.61	-35.72
Item 18	2.56	.73	-1.70	-35.53
Item 19	2.67	.63	-2.05	-45.93
Social Behavior	15.22	3.55	-1.54	-33.62
Academic Behavior	13.54	4.30	-0.93	-21.26
Emotional Behavior	18.31	3.24	-1.43	-32.19
Total Behavior	46.97	9.52	-1.18	-27.29

Note. SD = Standard Deviation.

Table 17

Relations to Other Variables (Spearman Correlation Coefficients)

SAEBRS-TRS Scale	Attendance	ODR: Minor	ODR: Major
Social Behavior	-.060*	-.186*	-.278*
Academic Behavior	-.159*	-.137*	-.211*
Emotional Behavior	-.065*	-.124*	-.204*
Total Behavior	-.123*	-.164*	-.253*

Note. ODR = Office Discipline Referral.

*$p < .05$

Table 18

Fit Statistics for Model Comparison of SAEBRS-TRS Scores

Model	χ^2	*df*	RMSEA	CFI	TLI	SRMR	Factor Loading Range
Unidimensional Model	7456.33*	152	.157	.883	.869	.131	TB: .560-.901
Second-Order Model	4518.60*	149	.123	.930	.920	.089	TB: .789-.930 SB: .832-.894 AB: .799-.931 EB: .647-.927
Bifactor Model	2227.52*	133	.090	.967	.957	.053	TB: .353-.895 SB: -.138-.526 AB: .310-.662 EB: .234-.843

Note. χ^2 = Chi-square goodness-of-fit test. *df* = degrees of freedom. RMSEA = Root Mean Square Error of Approximation. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. SRMR = Standardized Root Mean Square Residual. TB = Total Behavior. SB = Social Behavior. AB = Academic Behavior. EB = Emotional Behavior.

$^*p < .05$

Table 19

Fit Statistics of Single-Group Confirmatory Factor Analysis and Measurement Invariance Models

Model	*n*	χ^2 (*df*)	$\Delta\chi^2$ (*df*)	RMSEA	ΔRMSEA	CFI	ΔCFI	TLI	SRMR
Single-Group CFA									
Full Sample	1949	2227.52* (133)	-	.090	-	.967	-	.957	.053
Female	956	1078.59* (133)	-	.086	-	.964	-	.954	.063
Male	993	1318.87* (133)	-	.095	-	.968	-	.959	.052
Receives FRL	852	1135.13* (133)	-	.094	-	.967	-	.957	.057
Does not receive FRL	1097	1216.19* (133)	-	.086	-	.965	-	.956	.053
White	972	1053.53* (133)	-	.084	-	.968	-	.959	.052
Hispanic or Latino	355	593.54* (133)	-	.099	-	.963	-	.952	.080
Black	233	392.73* (133)	-	.092	-	.976	-	.970	.058
Asian	181	290.51* (133)	-	.081	-	.973	-	.966	.090
Two or More Races	170	291.88* (133)	-	.084	-	.977	-	.971	.058
Invariance (Sex Assigned at Birth)									
Configural	1949	2404.49* (300)	-	.085	-	.967	-	.962	.058
Metric/Scalar	1949	2862.03* (323)	441.97* (23)	.090	.005	.960	.007	.958	.062
Invariance (FRL Eligibility)									
Configural	1949	2467.96* (300)	-	.085	-	.965	-	.961	.058
Metric/Scalar	1949	2384.21* (323)	93.64* (23)	.081	.004	.967	.002	.965	.058
Invariance (Racial/Ethnic Identity)									
Configural	1911	2566.53* (801)	-	.076	-	.973	-	.971	.064
Metric/Scalar	1911	2786.40* (893)	438.38* (92)	.074	.002	.971	.002	.972	.067

Note. CFA = Confirmatory Factor Analysis. FRL = Free or Reduced-Price Lunch. x^2= Chi-square goodness-of-fit test. *df* = degrees of freedom. RMSEA = Root Mean Square Error of Approximation. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. SRMR = Standardized Root Mean Square Residual.

**p* < .05

Figure 2

Confirmatory Factor Analysis of SAEBRS-TRS Scores (Full Sample)

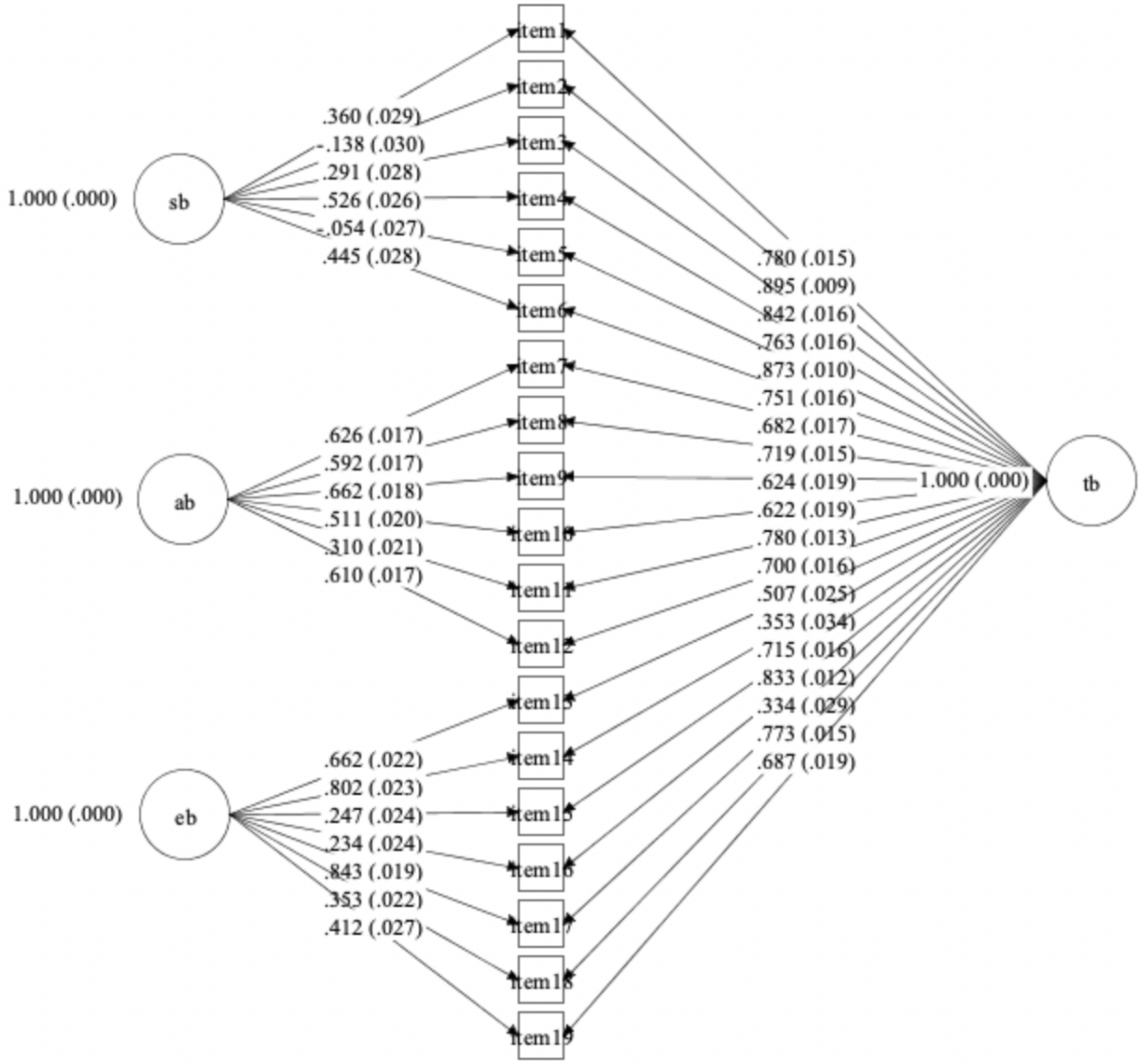


Figure 3

Confirmatory Factor Analysis of SAEBRS-TRS Scores (Female)

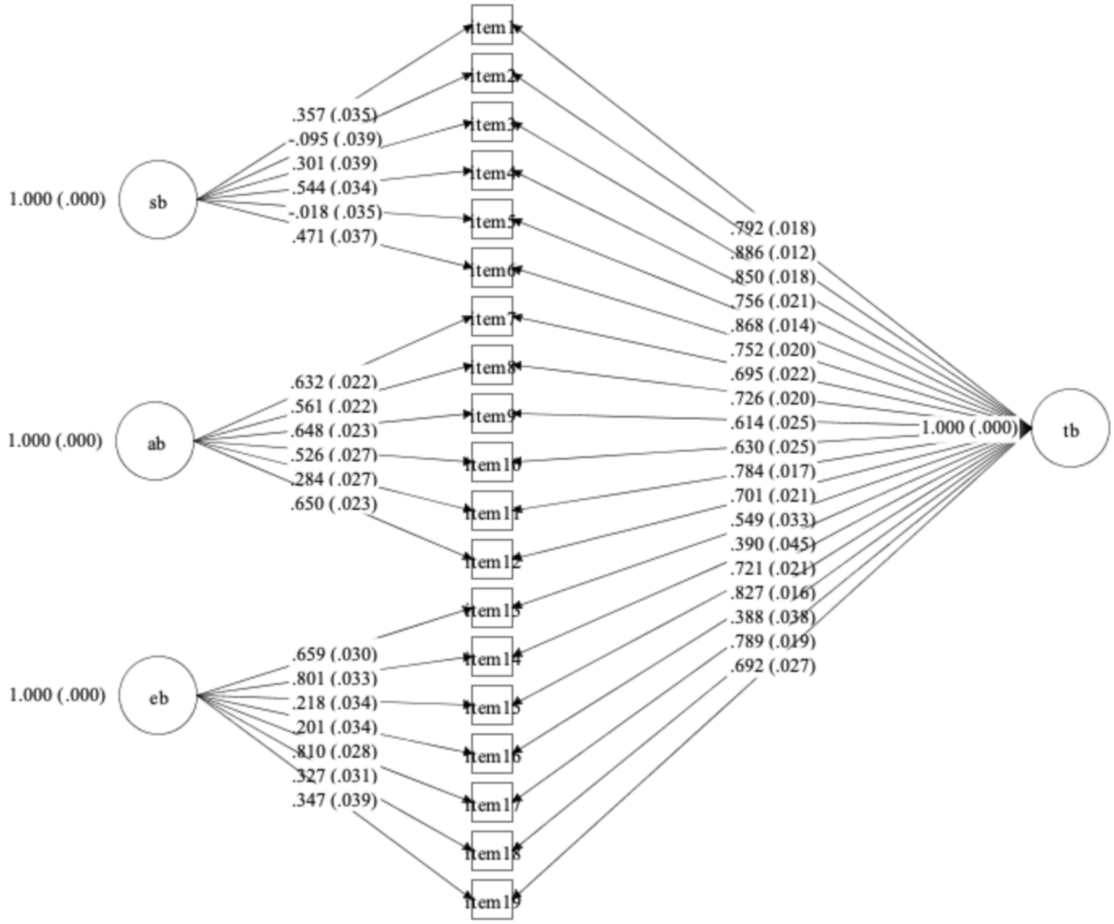


Figure 4

Confirmatory Factor Analysis of SAEBRS-TRS Scores (Male)

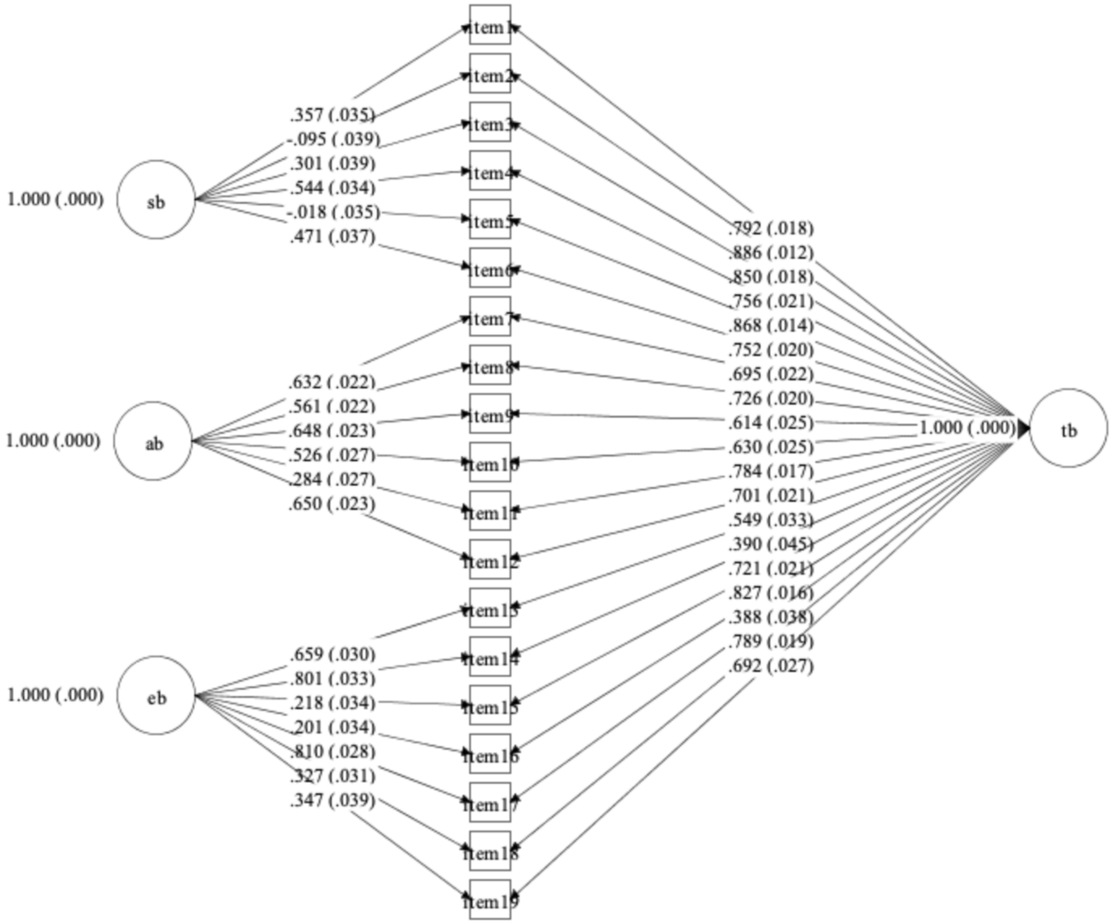


Figure 5

Confirmatory Factor Analysis of SAEBRS-TRS Scores (Eligible for Free or Reduced-Price Lunch)

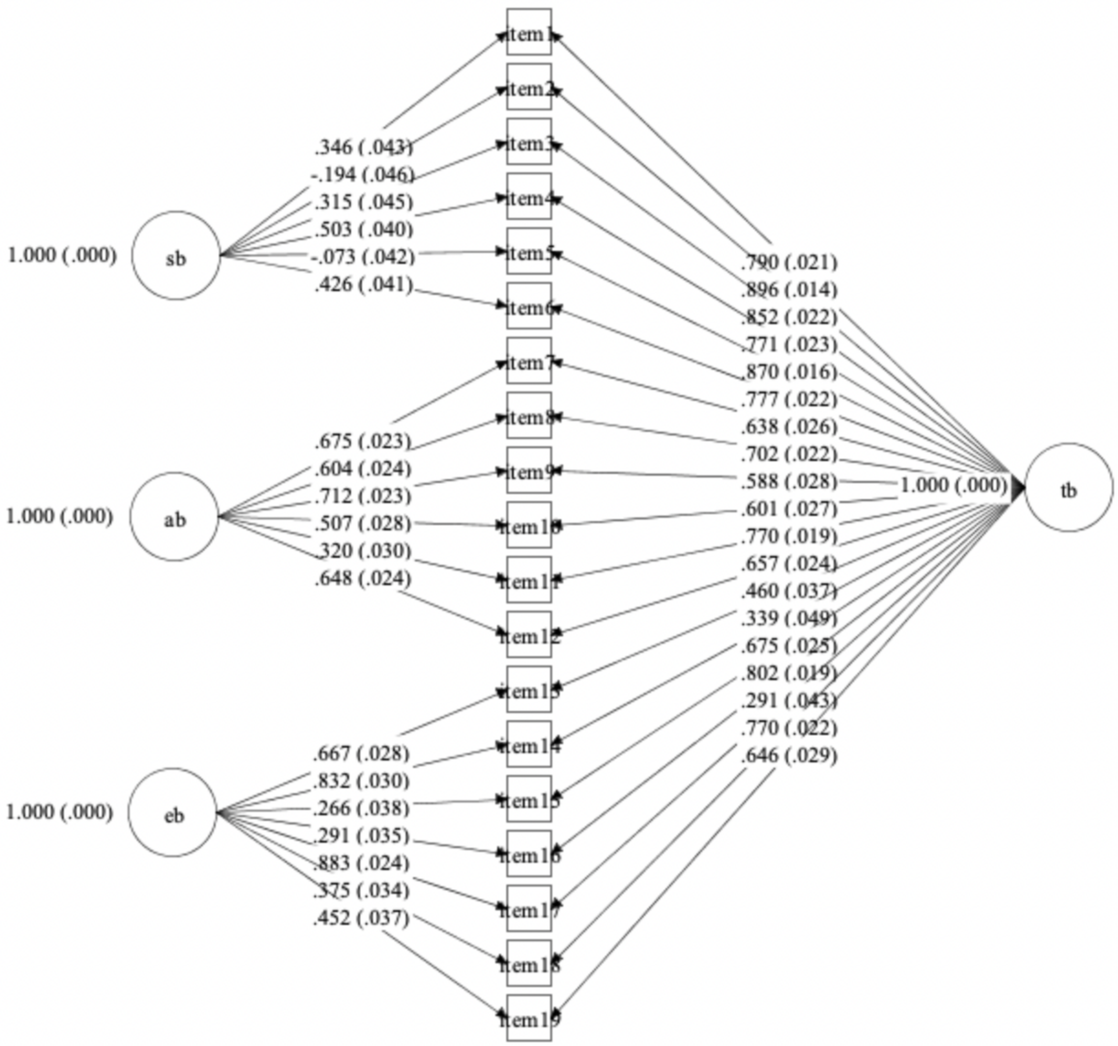


Figure 6

Confirmatory Factor Analysis of SAEBRS-TRS Scores (Not Eligible for Free or Reduced-Price Lunch)

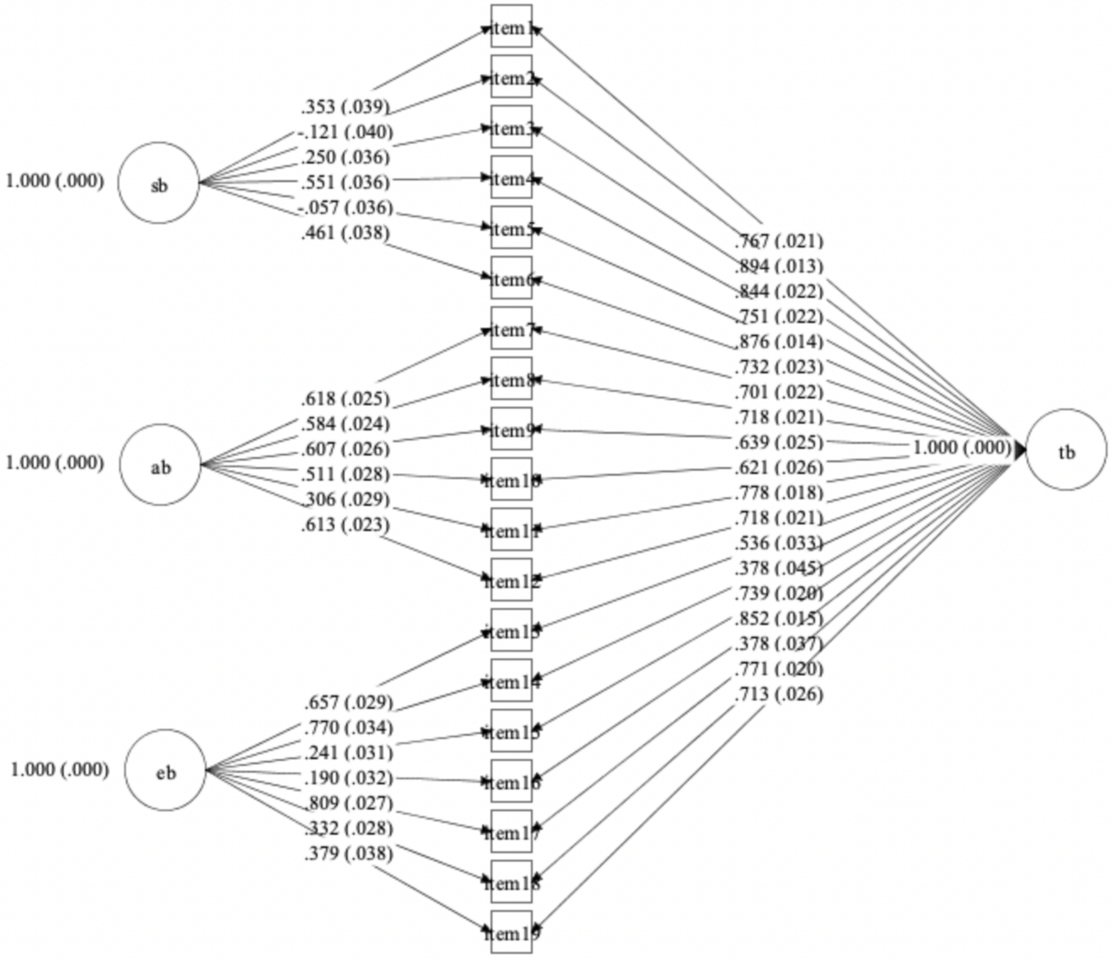


Figure 7

Confirmatory Factor Analysis of SAEBRS-TRS Scores (White)

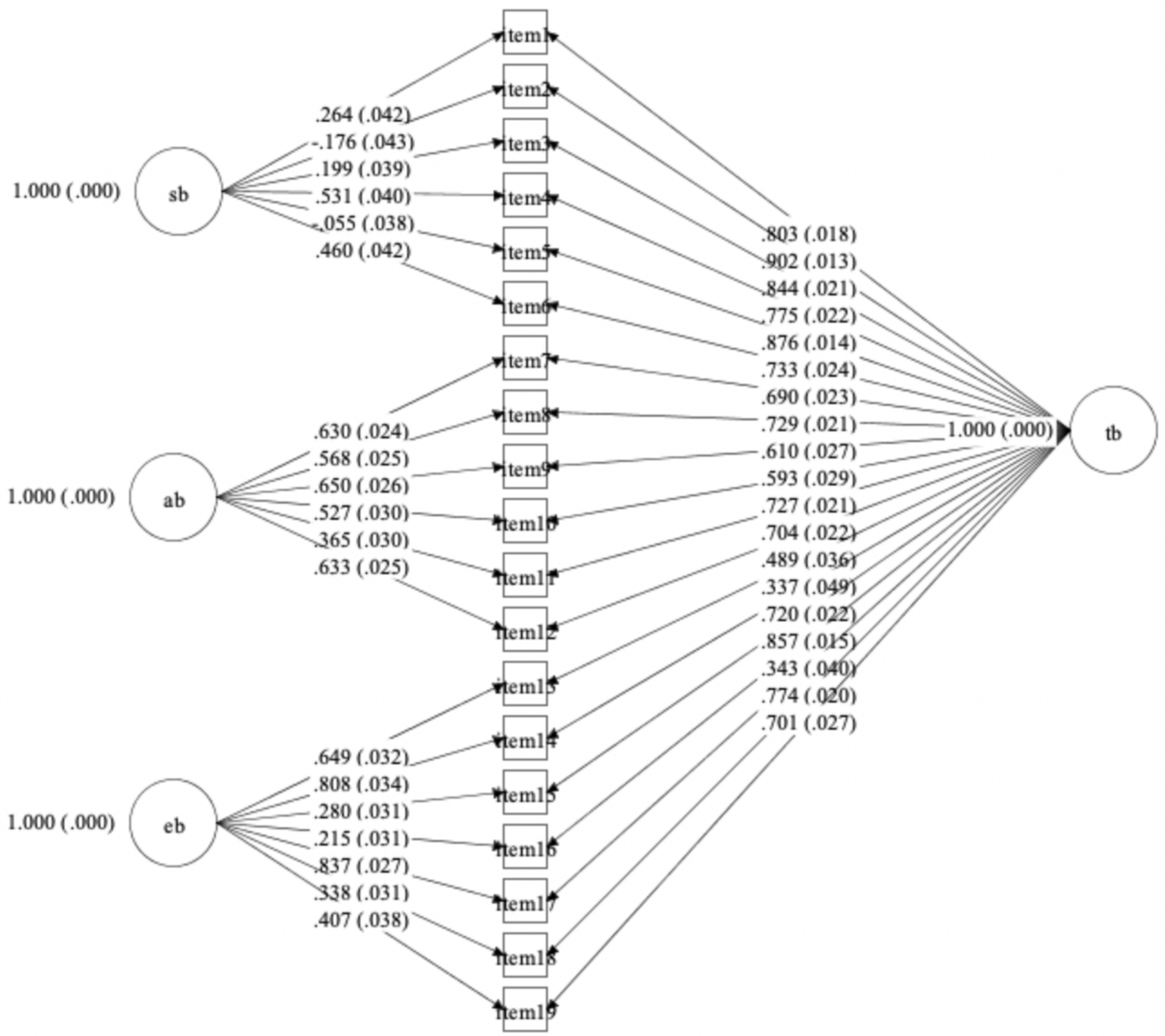


Figure 8

Confirmatory Factor Analysis of SAEBRS-TRS Scores (Black)

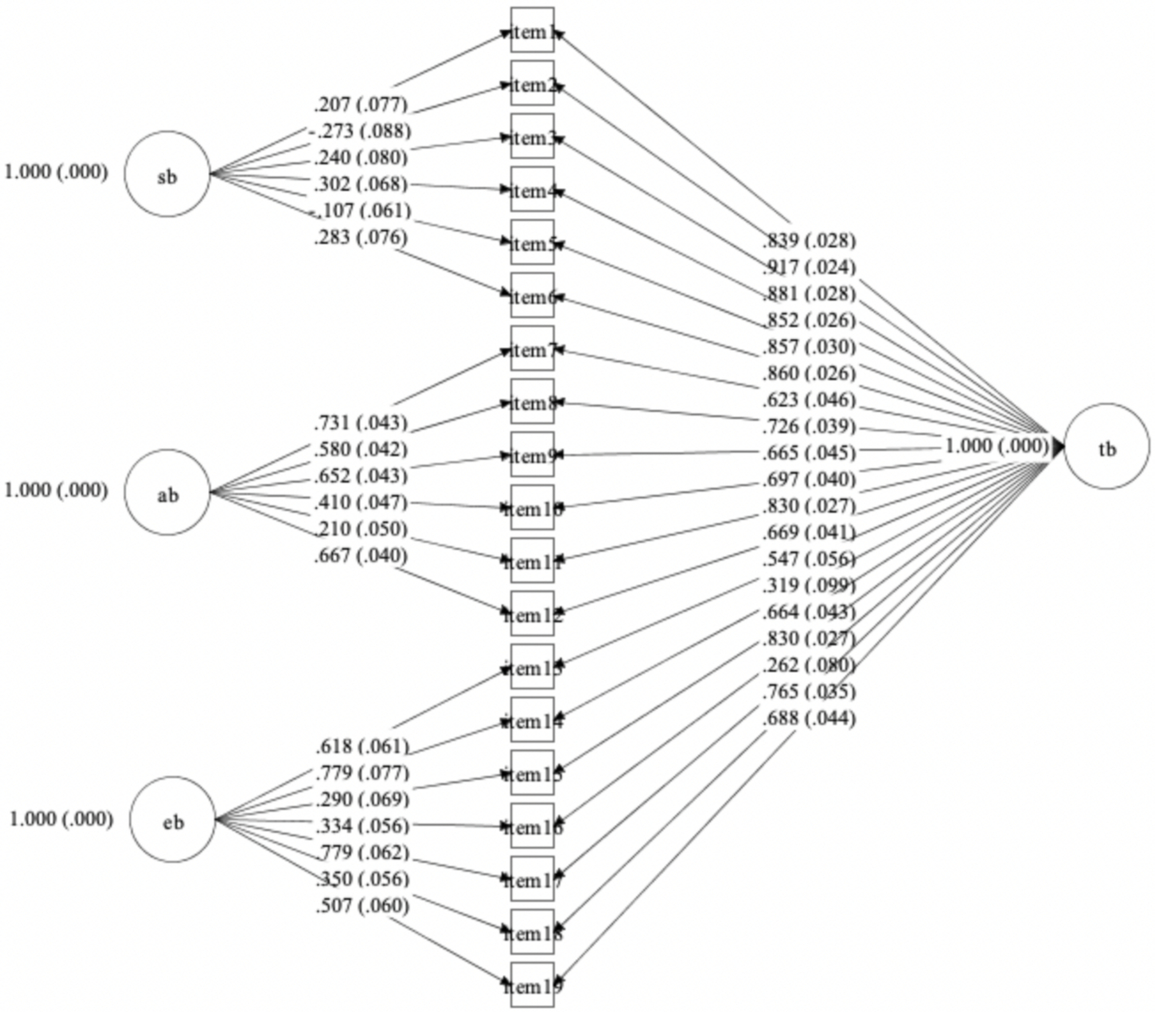


Figure 9

Confirmatory Factor Analysis of SAEBRS-TRS Scores (Asian)

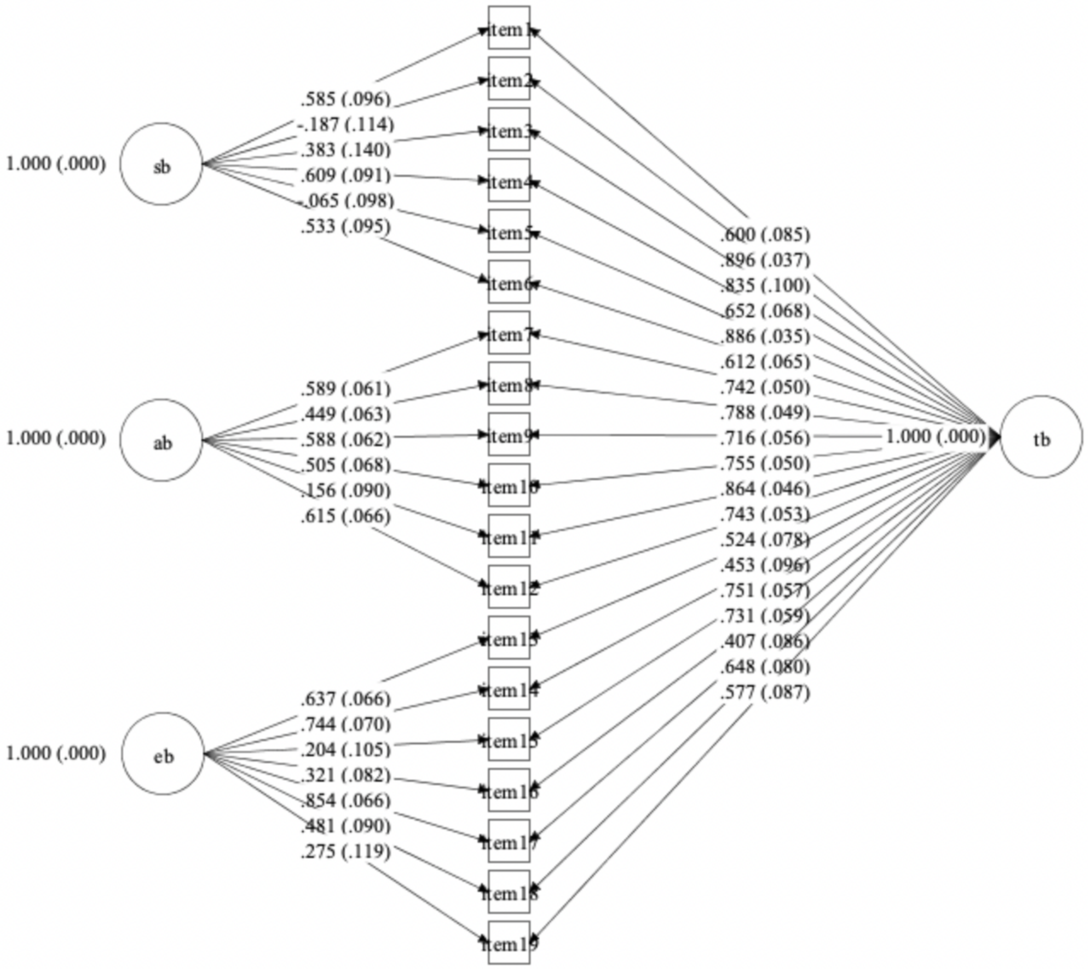


Figure 10

Confirmatory Factor Analysis of SAEBRS-TRS Scores (Two or More Races)

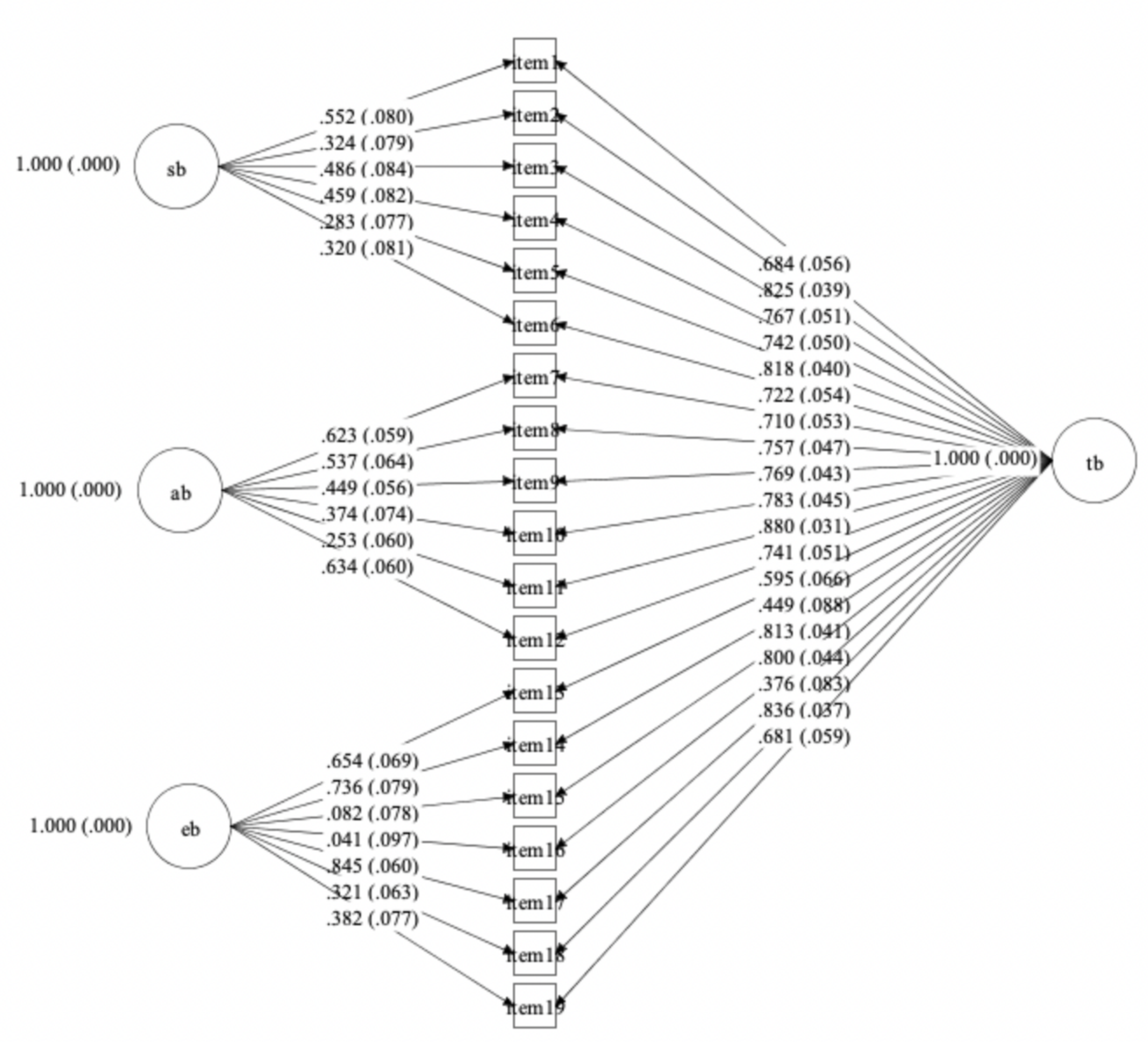
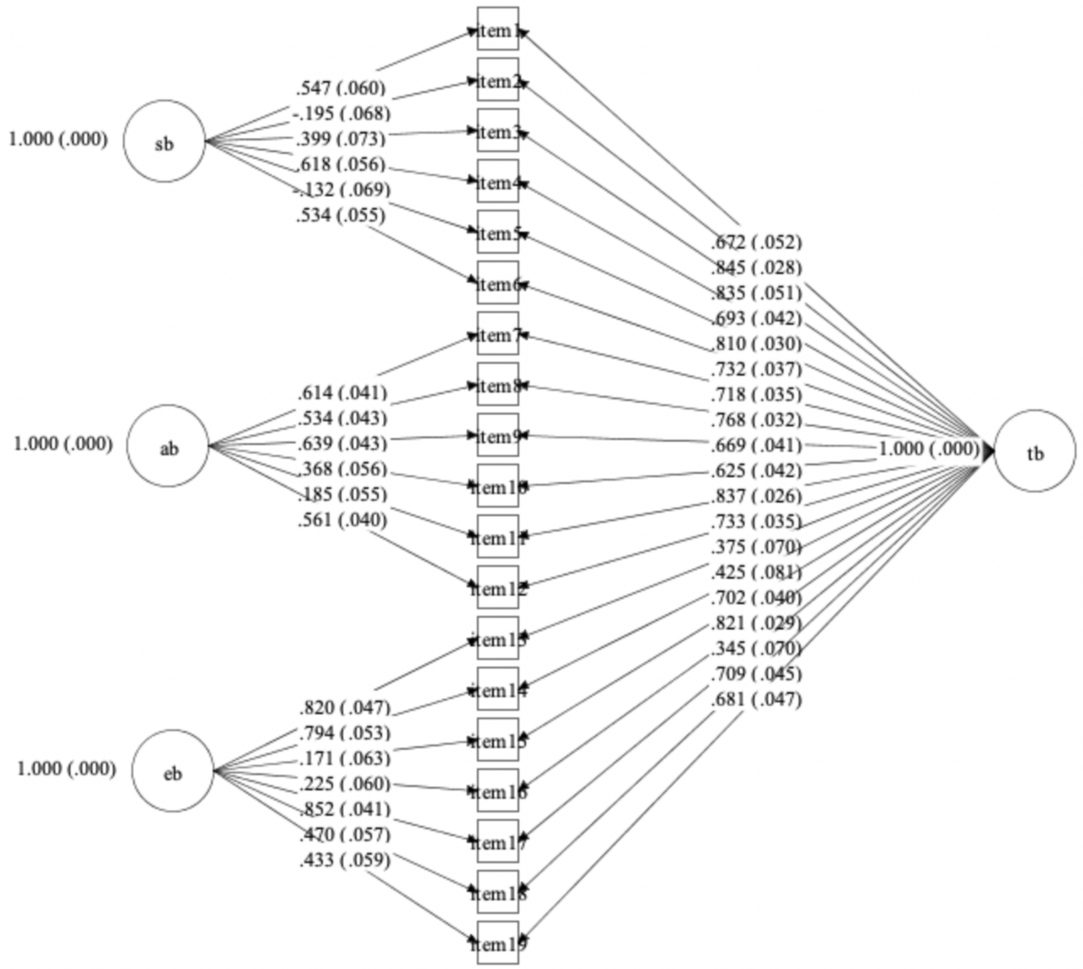


Figure 11

Confirmatory Factor Analysis of SAEBRS-TRS Scores (Hispanic or Latino)



Chapter 4

Synthesis & General Discussion

This multi-study dissertation aimed to examine the technical adequacy of the SAEBRS-TRS. Specifically, the first study synthesized extant literature for the SAEBRS-TRS in order to better understand sources of validity evidence to support the use and interpretation of scores. This synthesis revealed a gap in the SAEBRS-TRS literature related to evaluations of measurement bias; therefore, the second study investigated the potential invariance of the measure. The results of this dissertation provide preliminary evidence of the SAEBRS-TRS technical adequacy across sources of validity evidence. Yet, more research is needed to gather a stronger research base for using the SAEBRS-TRS with students from different backgrounds and lived experiences.

Implications for Research

A number of research-related themes emerged through this dissertation. First, a prominent finding was the promising, yet preliminary evidence for the technical adequacy of the SAEBRS-TRS. Despite evidence of validity across a number of domains, a variety of validity sources are in need of additional evidence. Specifically, research is needed to corroborate the promising findings at the elementary level in the geographic location of the Midwest. In particular, more research is needed to examine the psychometric properties of the SAEBRS-TRS at the secondary level and in locations outside of the Midwest. Further, future research must examine the consistency of the measure across raters, especially using the recent developments of the student version (mySAEBRS) and parent version (SAEBRS-PRS) of the measure. In addition, there is a clear need to investigate the negative factor loadings on the Social Behavior subscale within the bifactor model. Although items within a bifactor model often display a stronger association with the general factor, the negative and weak association with the specific factor is surprising and warrants further examination. Lastly, additional studies examining measurement bias (e.g., measurement invariance, differential item functioning) are critical for the SAEBRS-TRS.

Second, although evidence exists for the SAEBRS-TRS technical adequacy in relation to the populations of included studies, all findings may not be representative or generalizable to the broader population. Considering the growing cultural, linguistic, and economic diversity of the student population nationally, researchers must examine evidence of validity across populations. It is important for researchers to fill the gaps identified in this review for the SAEBRS-TRS, particularly related to the need for replication across a variety of student characteristics. Research also suggests that racial or ethnic alignment of students and teachers can have impacts on behavior ratings (Redding, 2019). Therefore, it is recommended that future research examine the racial or ethnic alignment of students and teachers using this screening tool and its impact on teacher ratings.

Third, there is a clear need for quality, comprehensive examinations of psychometric properties for *all* universal screening tools. Although this study examined one universal screener related to the SEB domains, a growing number of measures are available. Yet, the process of gathering validity evidence is time-consuming, inevitably leading to sources of validity with less evidence than other sources. In order for users to appropriately interpret findings with consideration of potential gaps, test developers should aim for transparency in directly reporting when there is lacking evidence related to any source of validity. Lastly, this dissertation suggests a need for researchers to utilize appropriate techniques for data analysis. Given the variety of available data analytic approaches and ever-changing best practices in measurement, it is recommended that researchers critically evaluate their data analytic approach and interpretation of analytical findings. Together, these studies offer guidance for future areas of research related to the SAEBRS-TRS.

Implications for Practice

Themes related to practice also emerged as a result of this dissertation. First, school psychologists are often regarded as data experts in educational settings. Yet, it can be time-consuming and difficult to critically review and evaluate the psychometric properties of universal screening tools. School psychologists, as well as other educators, must consult and collaborate to use their training and experience in measurement and assessment in order to provide recommendations and knowledge of best practices related to universal screening tools.

Second, practitioners should ensure they are engaging in appropriate use and interpretation of universal screening data. Although universal screening tools are not meant to be used in isolation for decision-making, educators may not have the data literacy skills to appropriately use and interpret data. Universal screening is promising for identifying students who may need additional support, yet the consequences of screening may be detrimental to students when interpreted incorrectly. Moreover, in consideration of the relatively weak findings for the EB subscale in comparison to the other scales, practitioners should avoid interpreting the subscales separately from the TB scale. As such, the use of multiple data points is recommended when making student-level decisions, such as intervention placement. This is important as universal screening by itself should lead to low-stakes decisions; however, it may lead to unintentional high-stakes decisions such as excluding students from intervention if their screening score results in a false-negative.

Practitioners may incorporate techniques such as data triangulation or a multiple gating procedure to make well-informed decisions about students. For example, a multiple gating procedure incorporates a comprehensive assessment process using multiple assessment tools, with each representing a “gate” for decision-making (Severson et al., 2007). While the process of using multiple data points to triangulate data takes more time and resources (i.e., multi-trait multi-method; Campbell & Fiske, 1959), it may result in more accurate identification of students (Severson et al., 2007).

Lastly, across both studies, a clear theme emerged regarding the generalizability and appropriateness of SAEBRS-TRS use across student characteristics. Practitioners are encouraged to examine the appropriateness of the SAEBRS-TRS and other screeners for use with their student population. Of importance, technical manuals may be available for practitioners to review and provide a brief overview of available research. Yet, technical manuals may also be used to sell a product to schools or school districts by emphasizing the positive findings of research. Comparatively, comprehensive syntheses of available literature, such as this dissertation, provide an unbiased overview of the SAEBRS-TRS technical adequacy to assist practitioners in reviewing the available literature when making decisions about universal screening tools.

Conclusion

This dissertation provides a thorough review of the technical adequacy of the SAEBRS-TRS, a widely used universal screening tool for the social, emotional, and behavioral domains. Although this synthesis provides supportive, preliminary evidence for the SAEBRS-TRS, researchers and practitioners are encouraged to critically evaluate the appropriateness for use with their intended population. Moreover, researchers and practitioners must interpret the data from the SAEBRS-TRS with a critical lens. Future research must prioritize minimizing the identified gaps in the literature.

References

American Educational Research Association, American Psychological Association, & National Center on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

American Psychological Association. (2020). P*ublication manual of the American Psychological Association 2020: The official guide to APA style* (7th ed.). American Psychological Association.

Arora, P. G., Sullivan, A. L., & Song, S. Y. (2022). On the imperative for reflexivity in school psychology scholarship, *School Psychology Review,* 1–13. https://doi.org/10.1080/2372966X.2022.2105050

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational & Psychological Measurement, 63*(1), 75–95. https://doi.org/10.1177/0013164402239318

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335–340. https://doi.org/10.3102/10769986027004335

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Bradley, R., Doolittle, J., & Bartolotta, R. (2008). Building on the data and adding to the discussion: The experiences and outcomes of students with emotional disturbance. *Journal of Behavioral Education, 17*(1), 4–23. https://doi.org/10.1007/s10864-007-9058-6

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford.

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

Bruhn, A. L., Woods-Groves, S., & Huddle, S. (2014). A preliminary investigation of emotional and behavioral screening practices in K–12 schools. *Education & Treatment of Children, 37*(4), 611–634. https://doi.org/10.1353/etc.2014.0039

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* Routledge.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

Castillo, W., & Gillborn, D. (2022). *How to "QuantCrit": Practices and questions for education data researchers and users*. Annenberg Institute at Brown University. https://doi.org/10.26300/v5kh-dd65

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Sage.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. Russell Sage Foundation.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology.* Harlow, UK: Pearson Education.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121–149. https://doi.org/10.1177/0748175610373459

Dowdy, E., Dever, B. V., DiStefano, C., & Chin, J. K. (2011). Screening for emotional and behavioral risk among students with limited English proficiency. *School Psychology Quarterly, 26*(1), 14–26. https://doi.org/10.1037/a0022072

Dowdy, E., Chin, J. K., & Quirk, M. P. (2013). Preschool screening: An examination of the Behavioral and Emotional Screening System Preschool Teacher Form (BESS Preschool). *Journal of Psychoeducational Assessment, 31*(6), 578–584. http://doi.org/10.1177/0734282913475779

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

Duong, M. T., Burns, E. J., Lee, K., Cox, S., Coifman, J., Mayworm, A., & Lyon, A. R. (2020). Rates of mental health service utilization by children and adolescents in schools and other common service settings: A systematic review and meta-analysis. *Administration and Policy in Mental Health Services Research, 48*(3), 420–439. https://doi.org/10.1007/s10488-020-01080-9

Edyburn, K. L., Dowdy, E., DiStefano, C., Bertone, A., & Greer F. (2020). Measurement invariance of the English and Spanish BASC-3 behavioral and emotional screening system parent preschool forms. *Early Childhood Research Quarterly, 51*(2), 307–316. https://doi.org/10.1016/j.ecresq.2019.12.002

*Eklund, K., Kilgus, S., von der Embse, N., Beardmore, M., & Tanner, N. (2016). Use of universal screening scores to predict distal academic and behavioral outcomes: A multilevel approach. *Psychological Assessment, 29*(5), 486–499. https://doi.org/10.1037/pas0000355

*Ezell, E. (2018). *Diagnostic accuracy of a universal behavior screener for identification of attention problems in first grade* [Unpublished master's thesis]. East Carolina University.

*Fallon, L. M., Veiga, M. B., Susilo, A., Robinson-Link, P., Berkman, T. S., Minami, T., & Kilgus, S. P. (2021). Exploring the relationship between teachers' perceptions of cultural responsiveness, student risk, and classroom behavior. *Psychology in the Schools, 59*(10), 1948–1964. https://doi.org/10.1002/pits.22568

Finney, S. J., & DiStefano, C. (2013). Dealing with nonnormality and categorical data in structural equation modeling. In G. R. Hancock, & R. O. Mueller (Eds.), *A second course in structural equation modeling* (2nd ed., pp. 439–482). Greenwich, CT: Information Age Publishing, Inc.

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology, 10*, 1507. https://doi.org/10.3389/fpsyg.2019.01507

Fitzpatrick, K. M., Drawve, G., & Harris, C. (2020). Facing new fears during the COVID-19 pandemic: The state of America's mental health. *Journal of Anxiety Disorders, 75,* 102291. https://doi.org/10.1016/j.janxdis.2020.102291

Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology, 3*(55), 1–21. https://doi.org/10.3389/fpsyg.2012.00055

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*(4), 625–641. https://doi.org/10.1080/10705510903203573

Fredrick, S. S., Drevon, D. D., & Jervinsky, M. (2019). Measurement invariance of the Student Risk Screening Scale across time and gender. *School Psychology, 34*(2), 159–167. https://doi.org/10.1037/spq0000278

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

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2006). Power and type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research, 41*(1), 29–53. https://doi.org/10.1207/s15327906mbr4101_3.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

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

*Hamsho, N. (2019). *Examining the classification accuracy of the Social, Academic, Emotional Behavior Risk Screener and its relationship with writing performance* [Unpublished doctoral dissertation]. Syracuse University.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Herman, K. C., Riley-Tillman, T., & Reinke, W. M. (2012). The role of assessment in a prevention science framework. *School Psychology Review, 41*, 306–314. https://doi.org/10.1080/02796015.2012.12087511

Higgins, J. P. T., & Thomson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Houri, A. K. (2020). *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* [Unpublished doctoral dissertation]. University of Minnesota.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118

*Iaccarino, S., von der Embse, N., & Kilgus, S. (2019). Interpretation and use of the Social, Academic, and Emotional Behavior Risk Screener: A latent transition approach. *Journal of Psychoeducational Assessment, 37*(4), 486–503. https://doi.org/10.1177/0734282918766650

Illuminate Education. (2021). *SAEBRS and mySAEBRS Norms and Benchmarks.* https://fastbridge.illuminateed.com/hc/en-us/articles/1260802344370

*Izumi, J. (2020). *Detecting and explaining differential item functioning on the Social, Academic, and Emotional Behavior Risk Screener* [Unpublished doctoral dissertation]. University of Missouri.

Jenkins, L. N., Demaray, M. K., Wren, N. S., Secord, S. M., Lyell, K. M., Magers, A. M., Setmeyer, A. J., Rodelo, C., Newcomb-Mcneal, E., & Tennant, J. (2014). A critical review of five commonly used social-emotional and behavioral screeners for elementary or secondary schools. *Contemporary School Psychology, 18*, 241–254. https://doi.org/10.1007/s40688-014-0026-6

Kalberg, J. R., Lane, K. L., & Menzies, H. M. (2010). Using systematic screening procedures to identify students who are nonresponsive to primary prevention efforts: integrating academic and behavioral measures. *Education and Treatment of Children, 33*, 561–584. https://doi.org/10.1353/etc.2010.0007

Kalberg, J. R., Lane, K. L., Driscoll, S., & Wehby, J. (2011). Systematic screening for emotional and behavioral disorders at the high school level: A formidable and necessary task. *Remedial and Special Education, 32*(6), 506–520. https://doi.org/10.1177/0741932510362508

Kamphaus, R. W., & Reynolds, C. R. (2007). *Behavior Assessment System for Children—Second Edition (BASC-2): Behavioral and Emotional Screening System (BESS).* Pearson.

Kamphaus, R. W., & Reynolds, C. R. (2015). *Behavior Assessment System for Children—Third Edition (BASC-3): Behavioral and Emotional Screening System (BESS).* Pearson.

*Kerry-Henkel, L. A. (2017). *Teacher burnout, self-efficacy, and the identification and referral of at-risk students* [Unpublished doctoral dissertation]. University of Arizona.

**Kilgus, S. P., Chafouleas, S. M., & Riley-Tillman, T. C. (2013). Development and initial validation of the Social and Academic Behavior Risk Screener for elementary grades. *School Psychology Quarterly, 28*(3), 210–226. https://doi.org/10.1037/spq0000024

**Kilgus, S. P., Sims, W. A., von der Embse, N. P., & Riley-Tillman, T. C. (2015). Confirmation of models for interpretation and use of the Social and Academic Behavior Risk Screener (SABRS). *School Psychology Quarterly, 30*, 335–352. https://doi.org/10.1037/spq0000087

*Kilgus, S. P., Eklund, K., von der Embse, N. P., Taylor, C. N., & Sims, W. A. (2016a). Psychometric defensibility of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) Teacher Rating Scale and multiple gating procedure within elementary and middle school samples. *Journal of School Psychology, 58*, 21–39. https://doi.org/10.1016/j.jsp.2016.07.001

*Kilgus, S. P., Sims, W. A., von der Embse, N. P., & Taylor, C. N. (2016b). Technical adequacy of the Social, Academic, and Emotional Behavior Risk Screener in an elementary sample. *Assessment for Effective Intervention, 42*, 46–59. https://doi.org/10.1177/1534508415623269

*Kilgus, S. P., Bowman, N. A., Christ, T. J., & Taylor, C. N. (2017). Predicting academics via behavior within an elementary sample: An evaluation of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS). *Psychology in the Schools, 54*, 246–260. https://doi.org/10.1002/pits.21995

*Kilgus, S. P., Bonifay, W. E., von der Embse, N. P., Allen, A. N., & Eklund, K. (2018a). Evidence for the interpretation of Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) scores: An argument-based approach to screener validation. *Journal of School Psychology, 68*, 129–141. https://doi.org/10.1016/j.jsp.2018.03.002

*Kilgus, S. P., Eklund, K., Maggin, D. M., Taylor, C. N., & Allen, A. N. (2018b). The Student Risk Screening Scale: A reliability and validity generalization meta-analysis. *Journal of Emotional and Behavioral Disorders, 26*(3), 143–155. https://doi.org/10.1177/1063426617710207

*Kilgus, S. P., von der Embse, N. P., Allen, A. N., Taylor, C. N., & Eklund, K. (2018c). Examining SAEBRS technical adequacy and the moderating influence of criterion type on cut score performance. *Remedial and Special Education, 39*(6), 377–388. https://doi.org/10.1177/0741932517748421

*Kilgus, S. P., von der Embse, N. P., Taylor, C. N., Van Wie, M. P., & Sims, W. A. (2018d). Diagnostic accuracy of a universal screening multiple gating procedure: A replication study. *School Psychology Quarterly, 33*(4), 582–589. https://doi.org/10.1037/spq0000246

*Kilgus, S. P., von der Embse, N. P., Eklund, K., Izumi, J. T., Van Wie, M. P., & Taylor, C. N. (2019). Co-occurrence of academic and behavioral risk within elementary schools: Implications for universal screening practices. *School Psychology, 34*(3), 261–280. https://doi.org/10.1037/spq0000314

*Kim, E., & von der Embse, N. (2021). Combined approach to multi-informant data using latent factors and latent classes: Trifactor mixture model. *Educational and Psychological Measurement, 81*(4), 728–755. https://doi.org/10.1177/0013164420973722

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). Guilford Press.

Leonardo, Z., & Grubb, W. N. (2019). *Education and racism: A primer on issues and dilemmas (2nd ed.).* Routledge.

Levitt, J. M., Saka, N., Romanelli, L. H., & Hoagwood, K. (2007). Early identification of mental health problems in schools: The status of instrumentation. *Journal of School Psychology, 45*, 163–191. https://doi.org/10.1016/j.jsp.2006.11.005

Lyon, A. R., & Bruns, E. J. (2019). From evidence to impact: Joining our best school mental health practices with our best implementation strategies. *School Mental Health, 11,* 106–114. https://doi.org/10.1007/s12310-018-09306-w

McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Erlbaum.

*McLean, D., Eklund, K., Kilgus, S. P., & Burns, M. K. (2019). Influence of teacher burnout and self-efficacy on teacher-related variance in social-emotional and behavioral screening scores. *School Psychology, 34*(5), 503–511. https://doi.org/10.1037/spq0000304

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568. https://doi.org/10.1037/0021-9010.93.3.568

Merikangas, K. R., He, J., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., Benjet, C., Georgiades, K., & Swendsen, J. (2010). Lifetime prevalence of mental disorders in U.S. adolescents: Results from the national comorbidity survey replication–adolescent supplement (NCS-a). *Journal of the American Academy of Child & Adolescent Psychiatry, 49*(10), 980–989. https://doi.org/10.1016/j.jaac.2010.05.017

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*(2), 184–196. https://doi.org/10.1037/spq0000085

Muthén, L. K., & Muthén, B. O. (2020). *Mplus user's guide* (8th ed.). Author.

Naser, S., Brown, J., & Verlenden, J. (2018). The utility of universal screening to guide school-based prevention initiatives: Comparison of office discipline referrals to standardized emotional and behavioral risk screening. *Contemporary School Psychology, 22*(4), 424–434. https://doi.org/10.1007/s40688-018-0173-2

Naser, S. C., & Dever, B. V. (2020). A preliminary investigation of the reliability and validity of the BESS-3 teacher and student forms. *Journal of Psychoeducational Assessment, 38*(2), 263–269. https://doi.org/10.1177/0734282919837825

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li., T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ: British Medical Journal, 372*(71), 1–9. https://doi.org/10.1136/bmj.n71

**Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology, 60*, 65–82. https://doi.org/10.1016/j.jsp.2016.11.002

*Pendergast, L. L., Youngstrom, E. A., Ruan-Iu, L., & Beysolow, D. (2018). The nomogram: A decision-making tool for practitioners using multitiered systems of support. *School Psychology Review, 47*(4), 345–359. https://doi.org/10.17105/SPR-2017-0097.V47-4

Perou, R., Bitsko, R. H., Blumberg, S. J., Pastor, P., Ghandour, R. M., Gfroerer, J. C., Hedden, S. L., Crosby, A. E., Visser, S. N., Schieve, L. A., Parks, S. E., Hall, J. E., Brody, D., Simile, C. M., Thompson, W. W., Baio, J., Avenevoli, S., Kogan, M. D., & Huang, L. N. (2013). Mental health surveillance among children--United States, 2005-2011. *MMWR supplements*, *62*(2), 1–35.

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*(3), 283–296. https://doi.org/10.7709/jnegroeducation.81.3.0283

Redding, C. (2019). A teacher like me: A review of the effect of student-teacher racial/ethnic matching on teacher perceptions of students and student academic and behavioral outcomes. *Review of Educational Research, 89*(4), 499–535. https://doi.org/10.3102/0034654319853545

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2012). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reynolds, C. R., & Suzuki, L. A. (2013). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology: Assessment psychology* (pp. 82–113). Wiley.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108–116. https://doi.org/10.7334/psicothema2013.260

*Roberson, A. J. (2019). *A longitudinal study of two teacher-report screening measures for student mental health: Comparing the SWTRS and SAEBRS* [Unpublished doctoral dissertation]. Louisiana State University.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. http://doi.org/10.1037/met0000045

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514. https://doi.org/10.1007/BF02296192

Severson, H. H., & Walker, H. M. (1992). *Systematic Screening for Behavior Disorders.* Sopris West.

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*(2), 193–223. https://doi.org/10.1016/j.jsp.2006.11.003

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Singh, S., Roy, D., Sinha, K., Parveen, S., Sharma, G., & Joshi, G. (2020). Impact of COVID-19 and lockdown on mental health of children and adolescents: A narrative review with recommendations. *Psychiatry Research, 293*, 113429. https://doi.org/10.1016/j.psychres.2020.113429

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *Urban Review, 34*, 317–342.

Skiba, R. J., Horner, R. H., Chung, C.-G., Rausch, M. K., May, S. L., & Tobin, T. (2011). Race is not neutral: A national investigation of African American and Latino disproportionality in school discipline. *School Psychology Review, 40*(1), 85–107. https://doi.org/10.1080/02796015.2011.12087730

Splett, J. W., Raborn, A., Lane, K. L., Binney, A. J., & Chafouleas, S. M. (2017). Factor analytic replication and model comparison of the BASC-2 Behavioral and Emotional Screening System. P*sychological Assessment, 29*(12), 1543–1549. https://doi.org/10.1037/pas0000458

Splett, J. W., Raborn, A., Brann, K., Smith-Millman, M. K., Halliday, C., & Weist, M. D. (2020). Between-teacher variance of students' teacher-rated risk for emotional, behavioral, and adaptive functioning. *Journal of School Psychology, 80*, 37–53. https://doi.org/10.1016/j.jsp.2020.04.001

*Stallone, D. (2021). *Comparing instrument efficacy in screening children with internalizing distress for a child-centered play therapy intervention* [Unpublished doctoral dissertation]. Roberts Wesleyan College.

Tang, S., Xiang, M., Cheung, T., & Xiang, Y. (2021). Mental health and its correlates among children and adolescents during COVID-19 school closure: The importance of parent-child discussion. *Journal of Affective Disorders, 279*(15), 353–360. https://doi.org/10.1016/j.jad.2020.10.016

*Tanner, N., Eklund, K., Kilgus, S. P., & Johnson, A. H. (2018). Generalizability of universal screening measures for behavioral and emotional risk. *School Psychology Review, 47*, 3–17. https://doi.org/10.17105/SPR-2017-0044.V47-1

*Taylor, C. N., Kilgus, S. P., & Huang, F. (2018). Treatment utility of universal screening for behavioral risk: A manipulated assessment study. *Journal of Applied School Psychology, 34*(3), 242–258. https://doi.org/10.1080/15377903.2017.1394949

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70. https://doi.org/10.1177/109442810031002

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

*von der Embse, N. P., Pendergast, L. L., Kilgus, S. P., & Eklund, K. R. (2016). Evaluating the applied use of a mental health screener: Structural validity of the Social, Academic, and Emotional Behavior Risk Screener. *Psychological Assessment, 28*(10), 1265–1275. https://doi.org/10.1037/pas0000253

*von der Embse, N. P., Kilgus, S. P., Ake, E., Eklund, K. R., & Levi-Nelson, S. (2018). Training teachers to facilitate early identification of mental and behavioral health risks. *School Psychology Review, 47*, 372–384. https://doi.org/10.17105/SPR-2017-0094.V47-4

*von der Embse, N. P., Kim, E. S., Kilgus, S., Dedrick, R., & Sanchez, A. (2019). Multi informant universal screening: Evaluation of rater, item, and construct variance using a trifactor model. *Journal of School Psychology, 77*, 52–66. https://doi.org/10.1016/j.jsp.2019.09.005

*Warmbold-Brann, K. (2017). *The effect of an intensive teacher training on the accuracy of social, emotional, and behavioral screening results* [Unpublished doctoral dissertation]. University of Missouri - Columbia.

*Weaver, E. S. (2019). *Progressive ratio and delay discounting assessments for young children with and without problem behavior* [Unpublished doctoral dissertation]. Vanderbilt University.

*Whitley, S. F., & Cuenca-Carlino, Y. (2019). Examining the technical adequacy of the Social, Academic, and Emotional Behavior Risk Screener. *Assessment for Effective Intervention, 46*(1), 67–75. https://doi.org/10.1177/1534508419857225

Whitney, D.G., & Peterson, M.D. (2019). US national and state-level prevalence of mental health disorders and disparities of mental health care use in children. *JAMA Pediatrics, 173*(4), 389–391. https://doi.org/10.1001/jamapediatrics.2018.5399

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*, 412–422. https://doi.org/10.1177/0741932508327463

*Youkhanna, V. W. (2021). *Do social, emotional, and behavioral problems impact vocabulary intervention?* [Unpublished doctoral dissertation]. University of California - Riverside.