Journal of School Psychology 58 (2016) 21–39

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

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☆

Stephen P. Kilgus [a,*], Katie Eklund [b], Nathaniel P. von der Embse [c], Crystal N. Taylor [a], Wesley A. Sims [a]

[a] *University of Missouri, United States*
[b] *University of Arizona, United States*
[c] *Temple University, United States*

ARTICLE INFO

Article history:
Received 25 May 2015
Received in revised form 16 April 2016
Accepted 1 July 2016
Available online xxxx

Keywords:
Universal screening
Behavioral and emotional risk
Teacher nomination
Rating scales

ABSTRACT

The primary purposes of this investigation were to (a) continue a line of research examining the psychometric defensibility of the Social, Academic, and Emotional Behavior Risk Screener — Teacher Rating Scale (SAEBRS-TRS), and (b) develop and preliminarily evaluate the diagnostic accuracy of a novel multiple gating procedure based on teacher nomination and the SAEBRS-TRS. Two studies were conducted with elementary and middle school student samples across two separate geographic locations. Study 1 ($n = 864$ students) results supported SAEBRS-TRS defensibility, revealing acceptable to optimal levels of internal consistency reliability, concurrent validity, and diagnostic accuracy. Findings were promising for a combined multiple gating procedure, which demonstrated acceptable levels of sensitivity and specificity. Study 2 ($n = 1534$ students), which replicated Study 1 procedures, further supported the SAEBRS-TRS' psychometric defensibility in terms of reliability, validity, and diagnostic accuracy. Despite the incorporation of revisions intended to promote sensitivity levels, the combined multiple gating procedure's diagnostic accuracy was similar to that found in Study 1. Taken together, results build upon prior research in support of the applied use of the SAEBRS-TRS, as well as justify future research regarding a SAEBRS-based multiple gating procedure. Implications for practice and study limitations are discussed.

1. Introduction

Schools are increasingly adopting multi-tiered systems of support (MTSS) as the foundation of their social–emotional and behavioral service delivery models (e.g., positive behavior interventions and supports; Bruhn, Woods-Groves, & Huddle, 2014). MTSS models represent a prevention-orientation to addressing student needs, with a foundation in ecological theory, data-based decision making, and problem solving logic (Burns, Riley-Tillman, & VanDerHeyden, 2012). Central to the application of MTSS is the use of evidence-based prevention and intervention strategies, which vary in level of intensity in the interest of supporting a wide range of students with varying levels of need. The application of these strategies is supported by the collection of assessment data, which inform a range of intervention-related decisions. One way to gather MTSS-relevant data is via universal screening (Chafouleas, Riley-Tillman, & Sugai, 2007), defined as the use of brief assessment tools to evaluate a population (e.g., all students within an elementary school) for the purpose of identifying individuals possessing some characteristic of interest (Jenkins, Hudson, & Johnson, 2007). Within MTSS models, these characteristics correspond to risk for social–emotional and behavioral

http://dx.doi.org/10.1016/j.jsp.2016.07.001
0022-4405/© 2016 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.

concerns, demonstrated via subsyndromal symptomatology predictive of future disordered behavior (Kamphaus, 2012). Documentation of such risk suggests a student may have not been exposed to universal prevention strategies, or rather is unresponsive to these strategies and will require more intensive intervention to improve their social–emotional and behavioral functioning.

Universal screening represents a key component of the MTSS process, supporting early identification of students who are at risk, and thus application of subsequent intervention and assessment practices (Cook, Volpe, & Livanis, 2010). Given its noted importance, universal screening for social–emotional and behavioral concerns has received a great deal of attention within the literature, with research yielding several screening tools. Multiple categories of such tools have been identified, including (a) multiple gating procedures, (b) evaluation of extant data collected as part of normal educational practices (e.g., office discipline referrals), and (c) teacher evaluation and rating of all students on common behavioral criteria (Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007). The majority of universal screening research includes this final category, representing brief rating scales comprised of a small number of Likert-scaled items (e.g., 12–30 items) that might be completed in only a few minutes.

A review of the literature reveals several examples of brief rating scale-based universal screening tools. These include the BASC-2 Behavioral and Emotional Screening System (BESS [25–30 items]; Kamphaus & Reynolds, 2007); the Social Skills Improvement System — Performance Screening Guide (SSIS-PSG [4 items]; Elliott & Gresham, 2008), the Strengths and Difficulties Questionnaire (SDQ [25 items]; Goodman, 1997); the Student Internalizing and Externalizing Behavior Screeners (SIBS/SEBS [14 items]; Cook, 2012; Cook et al., 2011); and the Student Risk Screening Scale—Internalizing and Externalizing (SRSS-IE [12 items]; Drummond, 1994). The majority of universal screeners utilize teacher rating scales, which teachers complete by rating the frequency with which students have exhibited a range of behavior during the past one or more months. Though promising, several limitations to existing screeners have been documented (Kilgus, Chafouleas, & Riley-Tillman, 2013), including (a) limited efficiency, with some screeners including a larger number of items (e.g., BESS, SDQ); (b) limited psychometric evidence, as some screeners are supported by a rather small number of studies (e.g., SIBS/SEBS, SSIS-PSG); and (c) a focus on problem behaviors alone, with some screeners not including content specific to the positive behaviors known to predict key students outcomes (Elias & Haynes, 2008; Kwon, Kim, & Sheridan, 2012). Recognition of these concerns has spurred development of novel screening tools (e.g., Daniels, Volpe, Fabiano, & Briesch, in press; Pennefather & Smolkowski, 2015). One brief teacher screener around which evidence has begun to accumulate is the Social, Academic, and Emotional Behavior Risk Screener — Teacher Rating Scale (SAEBRS-TRS; Kilgus, Chafouleas, Riley-Tillman, & von der Embse, 2014).

1.1. Social, Academic, and Emotional Behavior Risk Screener

The SAEBRS-TRS is a 19-item teacher rating scale available via FastBridge (fastbridge.org), an electronic web-based system of assessment tools. Specifically, the measure serves as the foundation of behavior universal screening within the system, and is a part of a broader suite of behavior assessment tools. When considered relative to Glover and Albers' (2007) evaluative criteria for universal screening tools, the SAEBRS-TRS is considered to afford several advantages. First, research to date supports SAEBRS-TRS technical adequacy, finding the measure to evidence several desirable psychometric properties. Several studies to date have supported SAEBRS-TRS reliability (Kilgus, Sims, von der Embse, & Riley-Tillman, 2015; Kilgus, Sims, von der Embse, & Taylor, in press; Kilgus et al., 2013; von der Embse, Pendergast, Kilgus, & Eklund, in press); concurrent validity, with findings supporting its capacity to predict the Social Skills Improvement System — Rating Scales (SSIS-RS; Gresham & Elliott, 2008), SRSS, and SIBS; and diagnostic accuracy (Kilgus et al., 2013, 2015; Kilgus, Sims, et al., in press). Interestingly, these latter studies have yielded somewhat inconsistent recommendations regarding which cut scores perform best within an applied SAEBRS classification model. Specifically, whereas two previous studies yielded similar cut score recommendations (relative to the SSIS-RS as a criterion; Kilgus et al., 2013, 2015), Kilgus, Sims, et al. (in press) identified a set of cut scores that differed from those selected via the previous two studies (relative to the SRSS and SIBS as criteria). These findings collectively suggest cut scores might vary in accordance with the outcome variable under consideration, thus rationalizing the need for additional research in this area.

Second, initial data have supported the SAEBRS-TRS' contextual appropriateness. Specifically, findings have suggested the SAEBRS-TRS might be used to predict student risk across multiple behavioral domains, including (a) Social Behavior (SB; 6 items), defined as behaviors that promote (e.g., social skills) or limit (e.g., externalizing problems) one's ability to maintain age appropriate relationships with peers and adults; (b) Academic Behavior (AB; 6 items), defined as behaviors that promote (e.g., academic enablers) or limit (e.g., attentional problems) one's ability to be prepared for, participate in, and benefit from academic instruction; and (c) Total Behavior (TB; 12 items), which incorporates all SB and AB items and is considered indicative of overall behavioral functioning. Von der Embse et al. (in press) expanded the comprehensiveness of the SAEBRS-TRS by introducing a new Emotional Behavior (EB; 7 items) subscale that is comprised of actions that promote (e.g., social–emotional competencies) or limit (e.g., internalizing problems) one's ability to regulate internal states, adapt to change, and respond to stressful/challenging events.

Third, initial SAEBRS-TRS research has yielded a brief screener considered to possess high usability, requiring a single rater to complete a small number of low effort suboperation actions (i.e., 19 Likert scale items). Such efficiency in instrumentation and proceduralization suggests the screener is likely to be highly acceptable to teachers, administrators, and other support staff (Kilgus et al., 2013, 2015). With that said, our experiences in the schools have also revealed general educator concerns regarding the efficiency of the overall screening process. A common question from teachers pertains to why it is necessary to screen all students, including (a) those they perceive as evidencing no risk of social–emotional or behavioral concerns, and (b) those they perceive as clearly evidencing risk for such concerns. Teachers contend that by ruling out these students and only evaluating those for whom risk status is uncertain, the universal screening process would become far more efficient.

1.2. Multiple gating procedures

Recognition of the above efficiency concerns in the past sparked initial interest in the development and evaluation of multiple gating procedures for universal screening (e.g., Walker, Severson, & Feil, 2014). Multiple gating procedures represent a broader assessment process incorporating two or more assessment methods, each of which represents a gate of the screening procedure with its own decision rules (Severson et al., 2007). Initial gates tend to be highly efficient, allowing for brief, low cost evaluation of all students. Subsequent gates corresponded to more comprehensive, time- and resource-intensive assessments that are only applied to those students identified as potentially at risk.

The primary goals of multiple gating procedures are to (a) streamline universal screening, resulting in an overall more efficient assessment process, and (b) increase the accuracy of screening decisions by corroborating data across multiple evidence-based methods (Severson et al., 2007). Research suggests the use of instruments in a multiple gate procedure is likely to increase diagnostic accuracy (Kilgus, Chafouleas, Riley-Tillman, & Welsh, 2012; Kilgus, Riley-Tillman, Chafouleas, Christ, & Welsh, 2014), with second gate measures helping to reduce false positives and increase true negatives by ruling out students incorrectly identified as at risk via an initial gate.

The *Systematic Screening for Behavior Disorders* (SSBD; Walker et al., 2014) is one example of a multiple gating procedure, wherein all students are universally evaluated using a highly efficient Gate 1 teacher nomination procedure. Nominated students are then evaluated via a Gate 2 behavior checklist and rating scale. Students found to exceed cut scores pass through to Gate 3, where they are examined via direct observation. Students exceeding cut scores on these observations are considered to be at risk and are referred for intervention. The SSBD has been described as a gold standard screening procedure (Lane et al., 2009; Severson et al., 2007), with a rich psychometric literature base supporting the procedure's defensibility (e.g., Caldarella, Young, Richardson, Young, & Young, 2008; Walker, Severson, Nicholson, & Kehle, 1994). A recent survey found that 14% of schools using universal screening procedures continue to use the SSBD (Bruhn et al., 2014). This wide dissemination, as well as its documented psychometric defensibility, suggests the SSBD might serve as a viable model for constructing highly usable and efficient universal screening procedures.

1.3. Summary & purpose

The current study was intended to address two overarching purposes. The first purpose was to continue validation efforts relative to the SAEBRS-TRS scales. The second purpose was to examine whether SAEBRS-TRS efficiency might be enhanced by embedding it within a multiple gating procedure akin to the SSBD. A proposed structure for such a SAEBRS-based multiple gating procedure is as follows. At Gate 1, teachers complete a systematic teacher nomination procedure intended to identify students fitting operational definitions for behavioral risk. At Gate 2, teachers complete the SAEBRS-TRS for those students nominated within Gate 1. Students found to exceed cut scores are considered at risk and in need of interventions and supports.

Six research questions were of interest. First, what is the internal consistency reliability associated with each SAEBRS-TRS scale? It was hypothesized the SAEBRS-TRS scales would yield the reliability required to supported low stakes decisions (>.80; Cortina, 1993; Nunnally, 1978). Second, what is the concurrent validity associated with each SAEBRS-TRS scale, as compared to the BESS? It was hypothesized SAEBRS-TRS scales would be moderately to highly associated with the BESS, a commonly used behavior screener with a rich psychometric history (Kilgus, Riley-Tillman et al., 2014), which was considered a criterion measure for the purposes of this investigation. Third, what is the unique contribution of each SAEBRS subscale relative to the prediction of behavioral and emotional risk? It was hypothesized that though the subscales would be correlated with each other, they would still each uniquely predict student risk per the BESS.

Fourth, what is the diagnostic accuracy associated with each scale of the SAEBRS-TRS? It was anticipated each scale would yield moderate to high overall diagnostic accuracy (as evaluated via the area under the curve [AUC] statistic), as well as cut scores associated with acceptable-optimal conditional probability values (i.e., sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio, and negative likelihood ratio). It was further anticipated that the SB and AB cut scores identified as most appropriate via previous investigations (i.e., Kilgus et al., 2013, 2015) would also perform well within the current study. This was given similarities between the current and prior studies in terms of the constructs assessed via the gold standard outcome measures (e.g., social skills, externalizing problems, attention problems).

Fifth, what is the diagnostic accuracy associated with the SAEBRS-based teacher nomination tool (SAEBRS-TN; i.e., Gate 1 of the proposed SAEBRS-based multiple gating procedure)? Although not intended for use in standalone fashion, it is necessary for the teacher nomination to yield certain psychometric properties to be defensible as a first gate. Specifically, it is necessary for the nomination procedure to yield high sensitivity (i.e., true positive rate) and moderate–high specificity (i.e., true negative rate). This is given that although subsequent gates permit the exclusion of students who are truly not at risk and who were incorrectly identified at Gate 1 (and thus increases in specificity), they do not permit the inclusion of students who are truly at risk but who were not identified at Gate 1 (and thus increases in sensitivity). As such, although initial specificity may be below acceptable standards (.70; Hintze & Silberglitt, 2005; Kilgus et al., 2013), sensitivity should likely meet or exceed standards for optimal performance (.90; Jenkins et al., 2007; Kilgus et al., 2013) if SAEBRS-TN is to be viable as a first gate tool. Sixth, what is the diagnostic accuracy associated with the SAEBRS-based multiple gating procedure? It was hypothesized the multiple gating procedure would yield acceptable to optimal conditional probability values that either met or exceeded those resulting from the SAEBRS-TRS.

The current research consisted of two separate and independent studies, with Study 1 conducted with a southeastern sample, and Study 2 conducted with a southwestern sample. Furthermore, Study 1 and Study 2 were conducted sequentially, such that the results of Study 1 (conducted in January–February 2014) informed the methods and procedures associated with Study 2 (conducted in September–October 2014). Although Study 2's methods largely replicated those from Study 1, a few specific changes were made for Study 2 to support examination of hypotheses related to the enhancement of the SAEBRS-based multiple gating procedure's diagnostic accuracy; these changes, as well as justification for the two-study structure, are discussed in more detail below within Study 1's Discussion below.

2. Study 1

2.1. Method

2.1.1. Participants

Elementary ($n = 567$) and middle ($n = 297$) school students from the southeastern United States were enrolled in Study 1. Students were sampled from a K–5 elementary school and a grade 6–8 middle school that were both from the same rural school district. Across both schools, all 34 elementary teachers and 12 middle school teachers were asked to participate. Of these invited teachers, 100% agreed to participate. Table 1 contains student demographic information for participants from both schools.

2.1.2. Measures

As part of this investigation, teachers completed two brief screening tools (i.e., BESS and SAEBRS-TRS) with regard to each individual student participant in their classroom. Teachers also completed a SAEBRS-based teacher nomination form once for their entire classroom. Each of these measures is described in more detail below.

2.1.2.1. Social, Academic, and Emotional Behavior Risk Screener. The Social, Academic, and Emotional Behavior Risk Screener — Teacher Rating Scale (SAEBRS-TRS) is a universal screening measure used to detect student risk for behavioral concerns. Completion of the SAEBRS-TRS yields four scores, including three subscale scores (i.e., Social Behavior, Academic Behavior, and Emotional Behavior), and a broad scale score inclusive of items from all three subscales (i.e., Total Behavior [TB]). Previous research has suggested the SAEBRS-TRS is founded upon a latent bifactor structure, wherein the TB scale accounts for covariance among all items, and the three subscales account for residual variance within item clusters after controlling for TB (Kilgus et al., 2015; von der Embse et al., in press). This bifactor structure corresponds to a SAEBRS-TRS theoretical framework. The framework posits that student behavior may be subdivided into three categories. Social Behavior (SB; 6 items) is defined as behaviors that promote (e.g., social skills) or limit (e.g., externalizing problems) one's ability to maintain age appropriate relationships with peers and adults. Academic Behavior (AB; 6 items) is defined as behaviors that promote (e.g., academic enablers) or limit (e.g., attentional problems) one's ability to be prepared for, participate in, and benefit from academic instruction. Finally, Emotional Behavior (EB; 7 items) is defined as actions that promote (e.g., social–emotional competencies) or limit (e.g., internalizing problems) one's ability to regulate internal states, adapt to change, and respond to stressful/challenging events.

The SAEBRS-TRS uses a 4-point Likert scale (*Never*, *Sometimes*, *Often*, *Almost Always*), in which users are asked to rate "How frequently the student displayed each of the following behaviors during the previous month." Scores within each subscale are summed to yield a raw score for SB, AB, and EB. Subscale scores are then combined to produce the broad TB score. Scale scores are positively scaled, such that higher scores are considered indicative of more adaptive functioning and reduced risk for behavioral concerns.

Table 1
Demographic statistics for student participants.

	Elementary school		Middle school	
	Study 1	Study 2	Study 1	Study 2
	N (%)		*N* (%)	
N	567	712	297	822
Gender				
Male	267 (47.1)	381 (53.5)	150 (50.5)	433 (52.7)
Female	300 (52.9)	324 (45.5)	147 (49.5)	386 (47.0)
Missing	–	7 (1.0)	–	3 (0.4)
Ethnicity				
White	284 (50.1)	371 (52.1)	128 (43.1)	538 (75.6)
Black or African American	195 (34.4)	20 (2.8)	128 (43.1)	36 (5.1)
Hispanic	64 (11.3)	190 (26.7)	30 (10.1)	194 (27.2)
American Asian	3 (0.5)	19 (2.7)	0 (0)	18 (2.5)
Native American	–	33 (4.6)	–	2 (0.3)
More than one	21 (3.7)	50 (7.0)	11 (3.7)	34 (4.8)
Missing	–	29 (4.1)	–	–
Free/reduced lunch	– (61.9)	– (28)	– (53.7)	– (34)

Previous research has supported the psychometric defensibility of the SAEBRS-TRS (Kilgus et al., 2013, 2015; Kilgus, Sims, et al., in press; von der Embse et al., in press), including its (a) internal consistency and inter-rater reliability; (b) concurrent validity relative to a series of behavioral outcomes (i.e., office discipline referrals and suspensions), academic outcomes (i.e., curriculum-based measurement and statewide achievement tests), and behavior rating scales (i.e., Student Risk Screening Scale [SRSS; Drummond, 1994], Student Internalizing Behavior Screener [SIBS; Cook et al., 2011], and Social Skills Improvement System – Rating Scales [SSIS-RS; Gresham & Elliott, 2008]); and (c) diagnostic accuracy relative to the SSIS-RS, SIBS, and SRSS. Though recommendations have been somewhat inconsistent, this preliminary research has also informed the selection of SAEBRS-TRS cut scores to be used in differentiating between students who are either at risk or not at risk for behavioral concerns. Across two studies with samples from different geographic regions, Kilgus et al. (2013) identified SB ≤ 12 and AB ≤ 9 as most optimal, whereas Kilgus, Kazmerski, et al. (in press) and Kilgus, Sims, et al. (in press) identified SB ≤ 15, AB ≤ 14, EB ≤ 16, and TB ≤ 42 or 47 (with TB score depending on the criterion construct of interest; i.e., internalizing vs. externalizing, respectively) as most optimal. Von der Embse et al. (in press), through the use of item response theory analyses and multilevel bifactor modeling, supported both the overall factor structure across student and teacher levels, as well as strong item level discriminative properties. Results supported the use of the SAEBRS as a screener in differentiating between individuals with low levels of risk from those with moderate and severe levels of risk.

2.1.2.2. Behavioral and Emotional Screening System. Teachers completed the Child/Adolescent form (ages 6–16) of the Behavior and Emotional Screening System (BESS) as part of this investigation. The Teacher BESS is a 27-item screening tool adapted from the Behavior Assessment System for Children, Second Edition (BASC-2; Reynolds & Kamphaus, 2004), with items pulled from each of the four composites (i.e., externalizing problems, internalizing problems, school problems, and adaptive skills) comprising the BASC-2. The BESS yields a single score considered indicative of behavioral and emotional risk. Numerous studies have collectively supported the reliability, validity, and diagnostic accuracy of the total score (see Lane, Menzies, Oakes, and Kalberg (2012) and Jenkins et al. (2014) for summaries of evidence supporting the BESS). Reliability evidence reported via the BESS technical manual is particularly strong, with internal consistency ranging from .96 to .97, and test–retest equaling .91. Validity evidence is similarly strong, with correlations ranging from .62 to .90 with the BASC-2, equaling .76 with the Teacher Report Form (TRF; Achenbach & Rescorla, 2001), and equaling .73 with the Conners Teacher Rating Scale (Conners, 1997). The extensive line of BESS research has resulted in the BESS frequently being utilized as a criterion outcome measure within previous diagnostic accuracy studies (e.g., Burke et al., 2012; Chafouleas et al., 2013; Johnson et al., 2016; Kilgus et al., 2012; Kilgus, Riley-Tillman, et al., 2014; Miller et al., 2015). It should be noted, however, that use of the BESS as a primary criterion measure has been identified as a potential shortcoming within several prior investigations, as there is need to explore more in-depth criterion measures and procedures (e.g., SSIS-RS and BASC-2). Nonetheless, the BESS was included as the criterion for these two studies for two reasons. First, the BESS is founded upon a rich psychometric literature base, with numerous studies supporting its use as an indicator of behavioral and emotional risk (see Lane et al., 2012). Second, relative to more comprehensive measures (e.g., BASC-2), use of the BESS increased the practicability of the current research. That is, by using a more efficient criterion measure, it was possible for the teacher participants to complete study measures for a larger number of students without undue time restraints to complete lengthier measures.

Completion of the BESS is estimated to take 5–10 min per student, and requires a rating for each item on a 4-point Likert scale from 0 to 3 (*Never*, *Sometimes*, *Often*, *Almost Always*). Once completed, the total raw score is calculated and then converted to a *T* score ($M = 50$, $SD = 10$). These *T* scores are scaled such that higher values correspond to more problematic behavior and less adaptive functioning. (Note: this scaling is the opposite of the SAEBRS-TRS, wherein higher scores are indicative of less problematic behavior and more adaptive functioning.) *T* scores are then compared to classification levels in determining a given student's risk status. Cut points related to risk have been developed to classify students into three classification levels, including *normal* (*T* score of 60 or below), *elevated* (*T* score from 61 to 70), and *extremely elevated* (*T* score of 71 or higher). The current study used a dichotomous scale (at risk/not at risk) to classify concern for problem behavior. A total T score of 61 or above, 1 standard deviation above the mean, was considered at risk. Scores are based on a normed sample representative of the United States population, and have been found to be a reliable and valid estimate of academic and behavioral risk for children ages 3–18. Within Study 1, 15.4% and 14.9% of elementary and middle school students, respectively, were identified as being at risk for behavioral and emotional concerns based on BESS teacher ratings.

2.1.2.3. SAEBRS Teacher Nomination. The SAEBRS Teacher Nomination (SAEBRS-TN) is a brief screening instrument specifically developed for Study 1 to serve as a first gate within the proposed SAEBRS-based multiple gating procedure. Within the current series of studies, teachers used the SAEBRS-TN to nominate students that they believe to be at risk for behavior problems. Teachers then completed the SAEBRS-TRS for all students in their classroom. To clarify, despite using the SAEBRS-TN, teachers still completed the SAEBRS-TRS for all students. This permitted the evaluation of both the novel SAEBRS-based multiple gating procedure, as well as the SAEBRS-TRS in isolation.

In an effort to mirror the SAEBRS-TRS, the SAEBRS-TN was developed to include three stages. In the first, teachers selected up to five students in their classroom that most closely fit the following definition for risk for social behavior problems: “student displays behaviors that limit his/her ability to maintain age appropriate relationships with peers and adults.” Examples of risk corresponded to eight maladaptive behaviors, including arguing, temper outbursts, and disruptive behavior. Non-examples of risk corresponded to seven adaptive behaviors, including cooperation with peers, joining of peer activities, and compliance with adult directions. In the second stage of the SAEBRS-TN, teachers selected up to five students in their classroom that most closely fit the following definition for risk for academic behavior problems: “student displays behaviors that limit his/her ability to be prepared for, participate in, and benefit from academic instruction”. Examples of risk corresponded to seven maladaptive behaviors, including distractedness, cheating, and disorganization of materials and assignments. Non-examples of risk corresponded to seven adaptive behaviors, including interest in academic topics, production of acceptable work, and academic engagement. In the third and final stage of the SAEBRS-TN, teachers selected up to five students in their classroom that most closely fit the following definition for risk for emotional behavior problems: “Student displays actions that limit his/her ability to regulate internal states, adapt to change, and respond to stressful/challenging events”. Examples of risk corresponded to eight maladaptive behaviors, including sadness, worry, and withdrawal. Non-examples of risk corresponded to seven adaptive behaviors, including emotional regulation, self-awareness, and happiness.

2.1.3. Procedure

Subsequent to school district approval for the investigation, researchers met with administrators from both schools to describe the study and evaluate interest in participation. Once both schools agreed to participate, the first author met with teachers at each school to provide information regarding the background and rationale for the study, as well as determine which teachers were willing to participate. In a subsequent meeting (approximately 15–20 min in length), the first author provided teachers with parental opt-out forms for all students in their classroom, and instructed teachers regarding how to distribute and collect opt-out forms in accordance with study procedures. In addition, the first author provided specific information regarding the structure and purpose of the SAEBRS-TRS, BESS, and SAEBRS-TN. Teachers were also instructed on how to complete each screening tool. Time was allotted at the end of the meeting for teacher questions regarding study purposes and procedures.

As noted above, students were recruited for the investigation via an opt-out procedure (approved by both university and school district Institutional Review Boards). Schools provided the parents of each student with an opt-out form that was to be signed and returned to the school if they did not want their child to participate in the study. Parents were given two weeks to return the forms prior to the start of screening procedures. However, if an opt-out form was returned after this time, any data that had been collected with regard to the student in question was destroyed. Overall, out of the 871 students between schools, only seven opt-out forms were returned, indicating a 99.20% participation rate.

Following the completion of this two-week period, teachers began to complete screening tools for the students in their classroom for whom an opt-out form was not returned. Elementary teachers rated all students in their primary classroom, whereas middle school teachers rated students enrolled in their homeroom. All screening tools were completed electronically via the online survey software, Qualtrics. To access the tools, teachers were provided with two electronic links. The first link provided access to a survey package that included the SAEBRS-TN, as well as a teacher demographic form. Each teacher completed this survey package once for the entire classroom. The second link provided access to a survey package comprised of a student demographic form, the BESS, and the SAEBRS-TRS. The teachers completed this survey package for each student in the classroom. Teachers were given three weeks to complete all measures for each participating student at their leisure.

The latter survey package comprised of the BESS and SAEBRS-TRS was structured so that the order in which the BESS and SAEBRS-TRS were presented to the teacher was randomly determined each time they accessed the survey. This allowed for counterbalancing of screener completion procedures and thus the ruling out of order effects. Teachers were not instructed to complete either of the survey packages in a particular order, nor was information collected regarding the order in which these were completed. As such, it was not possible to rule out ordering effects in regards to the sequencing of teacher nomination and brief teacher rating scales.

Subsequent to the screening process, schools received summary information regarding the results of the BESS. Only BESS data were provided given the measure's established defensibility in informing school-based decisions. Given the de-identified nature of the screening data, no student-level information was provided; rather, schools received data in aggregate form, corresponding to the percentage of students scoring in the normal, elevated, and extremely elevated levels via the BESS across the school and within each grade. School administrators and problem solving teams then used these data to inform decisions regarding ongoing school initiatives and the allocation of resources.

2.1.4. Data analyses

A series of analyses were conducted in evaluating each of the aforementioned research questions (RQs) pertaining to the psychometric defensibility of the SAEBRS-TRS, SAEBRS-TN, and SAEBRS-based multiple gating procedures. All analyses were conducted twice, separated across the elementary (grades K–5) and middle school (grades 6–8) samples. See below for a description of analyses as they pertain to each RQ. Note that all analyses were conducted using R Version 3.1.2.

2.1.4.1. RQ #1. The internal consistency reliability of each SAEBRS-TRS scale (i.e., SB, AB, EB, and TB) was evaluated via the calculation of Cronbach's alpha coefficients. Observed values were examined relative to previously recommended criteria for reliability required to support low-stakes decisions, including universal screening. In the current context, this criteria corresponded to Cronbach's alpha > .80 (Cortina, 1993; Nunnally, 1978).

2.1.4.2. RQ #2. Pearson product–moment bivariate correlation coefficients (r) were calculated to examine the concurrent criterion-related validity of each of the four SAEBRS-TRS scales. Specifically, each continuously-scaled SAEBRS-TRS score was compared to continuously-scaled BESS T scores. Resulting correlations were examined relative to behavior screening-specific evaluative criteria for correlational magnitude. Kilgus, Riley-Tillman, et al. (2014) established these criteria in their review of the extensive

Student Risk Screening Scale (Drummond, 1994) literature by synthesizing the available correlational evidence and noting the 25th ($r = .41$), 50th ($r = .58$), and 75th ($r = .69$) percentiles of correlation coefficients. These values were then used as criteria for small, moderate, and large correlations, respectively.

2.1.4.3. RQ #3. Two sets of analyses were used to examine the third research question. The first of these was correlation analyses, which examined the associated between the SAEBRS subscales. Moderate correlations were expected given (a) the presence of mono-method bias (i.e., all scores were derived from behavior rating scales), and (b) subscales were all measures of student behavior in the school setting (despite corresponding to different domains of behavior). The second set of analyses included logistic regressions, which examined the extent to which each SAEBRS subscale predicted student risk (per the BESS) after controlling for all other subscales. Statistics included the parameter estimates associated with each predictor, as well as the corresponding standard errors, z scores, and p values.

2.1.4.4. RQ #4. Diagnostic accuracy of the SAEBRS-TRS was examined via receiver operating characteristic (ROC) curve analysis, which was comprised of three steps. First, ROC analysis was used to calculate an area under the curve (AUC) statistic for each of the four SAEBRS-TRS scales. AUC is defined as the probability that a randomly selected student who is truly at risk (per the BESS within this investigation) would have a more at risk score on the SAEBRS relative to a randomly selected student who is not at risk. AUC has been described as an effect size-type statistic, indicative of the overall diagnostic accuracy associated with a scale (Swets, 1992). Values range between .50 and 1.00, with .50 corresponding to change or random decision making, and 1.00 to perfect decision making (i.e., 100% correct classification of students who are at risk and not at risk). Although recommendations have varied, commonly accepted recommendations for AUC interpretation call for values between .50 and .69 to be considered low, .70 and .89 moderate, and .90 and 1.00 high (Streiner & Cairney, 2007). Each AUC was evaluated via a test of statistical significance that examined the probability that the value was equal to .50; that is, the probability that the associated SAEBRS scale's diagnostic accuracy was equivalent to random decision making. AUC 95% confidence intervals (CIs) were also calculated using the DeLong, DeLong, and Clarke-Pearson (1988) asymptotic exact method.

Second, ROC analyses were used to simultaneously evaluate the performance of cut scores within each SAEBRS-TRS scale. Performance was expressed through six conditional probability statistics: (a) sensitivity, defined as the proportion of students who were truly at risk (per the BESS) who were found to be at risk via the SAEBRS-TRS; (b) specificity, defined as the proportion of students who were truly not at risk identified as not at-risk through the SAEBRS-TRS; (c) positive predictive value, or the proportion of students at risk on the SAEBRS-TRS who were truly at risk; (d) negative predictive value, or the proportion of students not at risk on the SAEBRS-TRS who were truly not at risk; (e) positive likelihood ratio, defined as how many times more likely an at-risk SAEBRS-TRS score was for students who were truly at risk (=Sensitivity / [1 − Specificity]); and (f) negative likelihood ratio, defined as how many times less likely a not at-risk SAEBRS-TRS score was for students who were truly at risk (=[1 − Sensitivity] / Specificity).

Two approaches to cut score evaluation were employed as part of this second step of ROC curve analysis, including calibration and cross-validation (Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008). The evaluation of EB and TB followed the calibration approach, wherein the performance of all possible cut scores was evaluated. This exploratory, post-hoc approach was chosen in recognition that only one prior study has yielded cut score recommendations for these scales. It is therefore necessary for subsequent research to both examine the performance of previously identified cut scores, while also considering whether alternative scores would be advisable. In contrast, the evaluation of SB and AB cut scores followed the cross-validation approach, such that the only cut scores evaluated within this study were those identified through multiple prior investigations as best suited for use in universal screening (i.e., SB ≤ 12 and AB ≤ 9). This verification-oriented approach has been described as a rigorous and necessary step in cut score evaluation, with Jenkins et al. (2007) noting "caution is warranted in using cut scores that have not been cross-validated across several samples of students" (p. 586).

Third, per the calibration approach, conditional probability statistics were reviewed to identify which EB and TB cut scores performed best and should therefore be selected as the basis of the SAEBRS risk classification model. Emphasis was placed on sensitivity and specificity within this process given the statistics' particular relevance to diagnostic accuracy and the evaluation of a universal screener's performance (Kilgus, Methe, Maggin, & Tomasula, 2014; Rutter & Gatsonis, 2001). The following evaluative criteria were used in evaluating cut score sensitivity: optimal ≥ .90, acceptable ≥ .80, and borderline ≥ .70 (Carran & Scott, 1992; Jenkins et al., 2007; Kilgus et al., 2013; Metz, 1978). Similar thresholds were used in the evaluation of specificity: optimal ≥ .80, acceptable ≥ .70, and borderline ≥ .60 (Hintze & Silberglitt, 2005). Cut score identification was founded upon the linear selection algorithm employed by Kilgus, Kazmerski, et al. (in press) and Kilgus, Sims, et al. (in press), which represented a revised version of that used within Kilgus et al. (2013) and Kilgus, Riley-Tillman, et al. (2014). The six steps were as follows: (1) optimal sensitivity and specificity, (2) acceptable sensitivity and optimal specificity, (3) optimal sensitivity and acceptable specificity, (4) acceptable sensitivity and specificity, (5) borderline sensitivity and acceptable specificity, and (6) acceptable sensitivity and borderline specificity. When examining the performance of all possible cut scores within each SAEBRS scale (e.g., SB), under ideal conditions, it would be possible to identify a cut score conforming to Step 1 criteria. If such a cut score was not available, an attempt was made to identify a cut score fitting Step 2 criteria. This process continued until a cut score was selected. If two or more cut scores were identified within the same step, the cut score selected was that which yielded the smallest difference between sensitivity and specificity. Once a cut score was selected, 95% CIs were calculated relative to its specific sensitivity and specificity values using a bootstrapping method with 2000 stratified bootstrapped replicates.

2.1.4.5. RQs #5 and #6. Four teacher nomination approaches and four multiple gating procedures were of interest (see Table 2 for a summary). Three of the four teacher nomination approaches corresponded to the three individual SAEBRS-TN forms (i.e., SB, AB, and EB). The fourth corresponded to a 'Combined' teacher nomination approach, wherein a student was considered to be at risk if they were nominated on one or more of the three individual SAEBRS-TN forms. Three of the four multiple gating procedures corresponded to pairings of construct-specific SAEBRS-TN forms (e.g., Social Behavior Problems) and SAEBRS-TRS subscale (e.g., Social Behavior [SB]). The remaining procedure corresponded to the pairing of the Combined nomination approach and the SAEBRS-TRS overall TB scale (which included all SAEBRS-TRS items).

Prior to the evaluation of their diagnostic accuracy, it was necessary to convert SAEBRS-TN and SAEBRS multiple gating procedure data into an analyzable scale. For each approach, this was done via the computation of dichotomous scores for each student. Within each of the four teacher nomination approaches, a student received a score of '1' if his or her teacher nominated the student as at risk, and a score of '0' if his or her teacher did not nominate the student. Within each of the four multiple gating procedures, a student received a score of '1' if both (a) his or her teacher nominated her as at risk and (b) the student received an at-risk score on the corresponding SAEBRS-TRS subscale (i.e., equal to or less than the selected cut score). In contrast, a student received a score of '0' for the multiple gating procedure if either (a) the student was not nominated by his or her teacher or (b) the student was nominated but did not receive an at-risk score on the corresponding SAEBRS-TRS subscale. Once dichotomous scores were calculated for each student across all four teacher nomination approaches and four SAEBRS-based multiple gating procedures, each approach was evaluated in terms of its diagnostic accuracy. This was done via comparing each of the eight scales to the dichotomously-scaled BESS risk scale (0 = not at risk and 1 = at risk), and then calculating the six conditional probability statistics described in the previous subsection.

2.2. Results

2.2.1. Missing data

Data were reviewed prior to analyses to identify the extent and nature of data missingness across the SAEBRS-TRS, SAEBRS-TN, and BESS measures. Overall, 29 student participants (3.36% of the sample) were missing data on one or more BESS or SAEBRS-TRS items. Results of Little's test suggested these data were missing completely at random (MCAR), $\chi^2 = 2584.45$, $p = .50$. In the interest of power retention, and thus maintaining the original sample of 864 students, missing data were handled via expectation–maximization, a single imputation technique. This particular approach was considered acceptable given the MCAR nature of the missing data (Enders, 2010).

Beyond SAEBRS-TRS and BESS data, a review of SAEBRS-TN data revealed nomination data were missing for 125 student participants (14.47% of the sample). Further review indicated data were missing as a result of teachers not completing the nomination survey package, which was separate from the BESS and SAEBRS-TRS survey package and accessed via a second electronic link. Recognition of the reason for nomination data missingness suggested SAEBRS-TN data were missing at random (MAR). The MAR assumption is considered tenable when it can be shown that missing data on variable *X* are predicted by one or more covariates but not by latent *X* values themselves (Enders, 2010). Given that nomination data were missing as a result of a teacher-level variable, and not as a result of a student-level variable, it was assumed missingness could not be predicted by student behavior (and thus dichotomous SAEBRS-TN scores), and data were thus MAR.

The decision was made to handle missing SAEBRS-TN data via listwise deletion. Though this approach typically requires the MCAR assumption, we felt use of listwise deletion was appropriate given previous simulation research, which indicated diagnostic accuracy findings are unlikely to differ meaningfully from "true" values when missing data rates are less than 30% (Janssen et al., 2010).

2.2.2. RQ #1

Table 3 contains the obtained reliability coefficients. At the elementary level, all Cronbach's alpha values were found to exceed the .80 threshold for low-stakes decision making (range = .83–.93; Cortina, 1993; Nunnally, 1978). The same was true of values at the middle school level, with the exception of EB, which yielded a Cronbach's alpha (.77) that approximated but did not meet the threshold. Across both grade levels, the broad TB scale was found to yield the highest Cronbach's alpha value (elementary = .93 and middle = .94), whereas the EB subscale yielded the lowest (elementary = .83 and middle = .77).

2.2.3. RQ #2

As can be seen in Table 3, findings were generally consistent across both grade levels, as all resulting correlation coefficients were in the expected direction, statistically significant at the $p < .001$ level, and met or exceeded the aforementioned universal screening-specific

Table 2
Social, Academic, and Emotional Behavior Risk Screener (SAEBRS)-based multiple gating procedures.

Multiple gating procedure type	Gate 1	Gate 2
Social	SAEBRS-TN Social	SAEBRS-TRS Social
Academic	SAEBRS-TN Academic	SAEBRS-TRS Academic
Emotional	SAEBRS-TN Emotional	SAEBRS-TRS Emotional
Combined	SAEBRS-TN Combined	SAEBRS-TRS Total

Table 3
Study 1 reliability, validity, and diagnostic accuracy coefficients.

Scale	Cronbach's alpha	Correlation with BESS	Area under the curve (AUC)
Elementary school			
Social Behavior	.89	−.79**	.90** (.86–.94)
Academic Behavior	.92	−.86**	.94** (.92–.97)
Emotional Behavior	.83	−.72**	.88** (.84–.92)
Total Behavior	.93	−.93**	.97** (.96–.99)
Middle school			
Social Behavior	.93	−.85**	.96** (.94–.97)
Academic Behavior	.92	−.88**	.95** (.92–.98)
Emotional Behavior	.77	−.69**	86** (.81–.91)
Total Behavior	.94	−.94**	.99** (.97–1.00)

Note. SAEBRS-TRS = Social, Academic, and Emotional Behavior Risk Screener; BESS = Behavioral and Emotional Screening System.
** $p < .001$.

threshold for high validity coefficient (i.e., .69). Of the SAEBRS scales, the broad TB scale yielded the highest correlations with the BESS (elementary = −.93 and middle = −.94), whereas the EB subscale yielded the lowest (elementary = −.72 and middle = .69).

2.2.4. RQ #3

Correlations between the SAEBRS subscales fell in the small to moderate ranges, with the correlation between SB and AB equaling .66, SB and EB equaling .61, and AB and EB equaling .56. Results of logistic regression analyses are presented in Table 4. At the elementary level, each subscale was found to be a statistically significant predictor of student risk after accounting for all other subscales. At the middle school level, only SB and AB were found to be statistically significant.

2.2.5. RQ #4

Across both grade levels, the overall diagnostic accuracy of each SAEBRS scale (per the AUC statistic) was statistically significant and in the moderate (.70–.90) or high (≥.90) range (see Table 3). The sole moderate AUC values corresponded to the EB subscale (i.e., elementary = .86 and middle = .88). Consistent with reliability and validity findings, the highest AUC values corresponded to the overall TB scale. At the elementary level, 95% CIs spanned from the moderate to high ranges for SB and EB, while entirely within the high range for AB and TB. Findings were similar at the middle school level, but with EB alone yielding CIs spanning the moderate and high ranges, and the remaining three scales' CIs lying within the high range only.

SAEBRS-TRS diagnostic accuracy was also supported via conditional probability statistics, which are included in Table 5. As noted above, the SB and AB scales were both evaluated via a confirmatory, cross-validation approach wherein the performance of a single set of cut scores was examined. These included SB ≤ 12 and AB ≤ 9, which were selected in consideration of two previous investigations (Kilgus et al., 2013, 2015). SB sensitivity was acceptable in elementary (>.80) and optimal (>.90) in middle school, whereas specificity was optimal across both grade levels (>.80). For AB, all sensitivity and specificity values fell in the optimal range across both elementary and middle school.

The exploratory calibration approach to cut score selection and evaluation resulted in the identification of adequately performing cut scores within the EB and TB scales. Findings were consistent, as the EB and TB scores selected at the elementary level were the same as those identified at the middle school level. Within EB, a cut score of 17 yielded optimal sensitivity and acceptable specificity at the elementary level, and acceptable sensitivity and specificity at the middle school level. Within TB, a cut score of 36 was associated with optimal sensitivity and specificity across both grade levels. See Table 6 for a summary of selected cut scores sensitivity and specificity values with associated 95% CIs.

Consistency was noted across all calibration and cross-validation cut scores in terms of (a) negative predictive values, which consistently approached 1.00, suggesting the majority of students identified as not at risk through SAEBRS scales were truly not at risk; and (b) comparatively lower positive predictive value (relative to negative predictive value), indicating that many

Table 4
Study 1 logistic regression analyses.

	Estimate	SE	z	p
Elementary school				
Intercept	12.72	1.68	7.57	<.001
SB	−0.33	0.07	−4.93	<.001
AB	−0.59	0.09	−6.72	<.001
EB	−0.30	0.07	−4.19	<.001
Middle school				
Intercept	12.95	3.00	4.32	<.001
SB	−0.59	0.12	−4.76	<.001
AB	−0.57	0.14	−4.15	<.001
EB	−0.19	0.14	−1.35	.179

Table 5
Study 1 conditional probability statistics associated with Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) teacher rating scales, teacher nomination approaches, and multiple gating procedure.

SAEBRS scale	Cut	SE	SP	PPV	NPV	LR+	LR−
Elementary school							
Social	**12**	**.81**	**.86**	**.51**	**.96**	**5.79**	**0.22**
Academic	**9**	**.91**	**.84**	**.52**	**.98**	**5.69**	**0.11**
Emotional	15	.67	.87	.49	.94	5.15	0.38
	16	.78	.80	.42	.95	3.90	0.28
	17	**.90**	**.73**	**.38**	**.97**	**3.33**	**0.14**
	18	.93	.61	.30	.98	2.38	0.11
Total	**36**	**.90**	**.93**	**.70**	**.98**	**12.86**	**0.11**
	37	.90	.90	.62	.98	9.00	0.11
	38	.92	.88	.58	.98	7.67	0.09
	39	.94	.85	.54	.99	6.27	0.07
Teacher Nomination	Social	.63	.91	.51	.93	7.00	0.41
	Academic	.66	.87	.50	.93	5.08	0.39
	Emotional	.36	.91	.43	.88	4.00	0.70
	Combined	.89	.76	.41	.97	3.71	0.14
MGP	Combined	.83	.95	.73	.97	16.60	0.18
Middle school							
Social	**12**	**.93**	**.85**	**.51**	**.99**	**6.20**	**0.08**
Academic	**9**	**.91**	**.83**	**.47**	**.98**	**5.35**	**0.11**
Emotional	15	.56	.90	.48	.92	5.60	0.49
	16	.79	.83	.45	.96	4.65	0.25
	17	**.86**	**.73**	**.35**	**.97**	**3.19**	**0.19**
	18	.95	.62	.30	.99	2.50	0.08
Total	**36**	**.95**	**.92**	**.66**	**.99**	**11.88**	**0.05**
	37	.95	.90	.61	.99	9.50	0.06
	38	.98	.89	.59	1.00	8.91	0.02
	39	1.00	.85	.52	1.00	6.67	0.00
Teacher Nomination	Social	.53	.92	.52	.93	6.63	0.51
	Academic	.50	.91	.45	.92	5.56	0.55
	Emotional	.28	.94	.42	.89	4.67	0.77
	Combined	0.82	0.79	0.40	0.96	3.90	0.18
MGP	Combined	0.79	0.95	0.72	0.96	15.80	0.22

Note. Bolded cut scores correspond to those selected within each SAEBRS-TRS subscale as yielding the most optimal performance. SE = sensitivity, SP = -specificity, PPV = positive predictive value, NPV = negative predictive value, LR+ = positive likelihood ratio, LR− = negative likelihood ratio, and MGP = -multiple gating procedure.

students identified as at risk through SAEBRS scales were actually not truly not at risk (e.g., 30–65%). Of the SAEBRS scales, TB was found to consistently perform best, yielding positive predictive value values that exceeded those of alternative scales. The opposite was true of the EB scale, with associated positive predictive values falling below those of other scales.

The magnitude of positive and negative likelihood ratios was evaluated relative to criteria outlined by DiCenso, Guyatt, and Ciliska (2014). Large positive likelihood ratios (>10) were noted for TB at both grade levels, moderate (5–10) for SB and AB at both grade levels, and small (2–5) for EB at both grade levels. In regards to negative likelihood ratios, large values (<0.1) were noted for SB and TB at the middle school level, moderate (0.1–0.2) for AB and EB at both grade levels and TB at the elementary level, and small (0.2–0.5) for SB at the elementary level.

Base rates were calculated for each SAEBRS-TRS to determine what percentage of students across the total sample would have been classified as at risk per the selected cut scores. Elementary level base rates were as follows: SB = 24.2%, AB = 26.5%, EB = 36.1%, and TB = 19.3%. Middle school level base rates were as follows: SB = 26.0%, AB = 28.9%, EB = 36.9%, and TB = 21.5%.

Table 6
Study 1 sensitivity (SE) and specificity (SP) values with corresponding bootstrapped 95% confidence intervals (CI-95) for selected SAEBRS-TRS cut scores.

SAEBRS-TRS scale	SE	SP	SE CI-95	SP CI-95
Elementary school				
Social Behavior	.81	.86	.73–.89	.83–.89
Academic Behavior	.91	.84	.85–.97	.81–.87
Emotional Behavior	.90	.73	.83–.96	.69–.77
Total Behavior	.90	.93	.83–.96	.90–.95
Middle school				
Social Behavior	.93	.85	.84–1.00	.81–.89
Academic Behavior	.91	.83	.81–.98	.78–.87
Emotional Behavior	.86	.73	.77–.95	.67–.78
Total Behavior	.95	.92	.88–1.00	.88–.95

2.2.6. RQ #5

Table 5 contains the conditional probability statistics associated with each of the four teacher nomination approaches (i.e., Social, Academic, Emotional, and Combined). On average, findings pertaining to the three individual forms were consistent in that although each yielded negative predictive values approaching 1.00 (.88–.93) and optimal specificity (.87–.94), each was also associated with less than acceptable levels of sensitivity (.28–.66). Positive predictive values were less than adequate, yet comparable to the values associated with the SAEBRS-TRS scale. Moderate positive likelihood ratios were noted for the Social and Academic forms at both grade levels, whereas small positive likelihood ratios were noted for EB at both grade levels. In regards to negative likelihood ratios, small values were noted for the Social and Academic forms at the elementary level. All of the remaining negative likelihood ratio values fell above the threshold for small values.

In contrast, the Combined nomination approach was found to yield adequate performance at both the elementary and middle school levels. Sensitivity and specificity both fell in acceptable ranges (i.e., ≥.80 and ≥.70, respectively), with positive and negative predictive values being similar to those associated with the aforementioned individual SAEBRS-TN forms. Positive likelihood ratios were found to be small (2–5), whereas negative likelihood ratios were moderate (0.1–0.2).

2.2.7. RQ #6

Given the unacceptable sensitivity associated with each SAEBRS-TN form, the choice was made to not move forward and examine the performance of the three multiple gating procedures associated with these individual forms. This was given that, as noted above, although multiple gating procedures allow one to improve upon the specificity of a first gate tool (i.e., by ruling out false positives via subsequent gates), one cannot improve upon a first gate's sensitivity (i.e., by reducing false negatives via subsequent gates). As such, if a student is incorrectly identified as not at risk at a first gate, he or she would not be included in latter gates. Yet, the performance of the Combined nomination approach supported its further consideration as part of a multiple gating procedure (see Table 5). The procedure was found to yield acceptable or nearly acceptable sensitivity at both school levels, optimal specificity, and positive and negative predictive values approximating those associated with the SAEBRS-TRS TB overall scale. Positive likelihood ratios fell in the large range, whereas negative likelihood ratios fell in the moderate range. To note, a review of resulting base rates suggested that if the Combined multiple gating procedure were to be used within the current sample, 17.3% and 16.2% of students would have been identified as at risk at the elementary and middle school levels, respectively.

2.3. Discussion

The purposes of Study 1 were twofold. The first was to further validate each of the four SAEBRS-TRS scales (i.e., SB, AB, EB, and TB). Results were promising, as scales were predominantly associated with acceptable internal consistency reliability (RQ #1), high concurrent criterion-related validity (RQ #2), and moderate to high diagnostic accuracy (RQ #4). Each subscale also demonstrated its unique contribution to the prediction of risk, with logistic regression analyses suggesting that with the exception of EB at the middle school level, each subscale was a statistically significant predictor of student risk after accounting for all other subscales (RQ #3).

Each scale also yielded cut scores associated with acceptable to optimal sensitivity and specificity. Cut scores of SB ≤ 12 and AB ≤ 9, which were identified via two prior studies and presently examined through a cross-validation approach, were associated with acceptable conditional probability statistics across both elementary and middle school grade levels. Through an exploratory calibration approach, cut scores of EB ≤ 17 and TB ≤ 36 were found to yield superior performance relative to alternative cut scores and were therefore selected within both elementary and middle school samples. That all four cut scores performed well within both samples speaks to their robustness as indicators of risk. Such continued cross-validation represents an important and necessary step toward justification for the use of the SAEBRS-TRS in universal screening within applied settings (e.g., schools).

The second purpose of Study 1 was to examine the diagnostic accuracy of SAEBRS-based multiple gating procedures. As noted above, the performance associated with the three individual SAEBRS-TN forms did not support their further consideration as part of multiple gating procedures (RQ #5). Yet, in contrast, the Combined nomination approach's adequate performance supported further evaluation. The resulting multiple gating procedure, which integrated the Combined nomination approach (Gate 1) with the SAEBRS-TRS TB overall scale (Gate 2), performed adequately at both the elementary and middle school levels (RQ #6). Specifically, the Combined multiple gating procedure performed similarly to the SAEBRS-TRS TB scale (the best performing individual SAEBRS-TRS scale) in terms of five of six conditional probability statistics (i.e., specificity, positive predictive value, negative predictive value, positive likelihood ratio, and negative likelihood ratio). An exception in performance pertained to sensitivity, with the Combined multiple gating procedure yielding values that fell within or just below the acceptable range (≥.80).

Taken together, these latter diagnostic accuracy findings call into question the use of the three individual SAEBRS-TN forms. However, results appeared to support the Combined nomination approach, as well as the combined multiple gating procedure. Though the procedure's performance did not meet the standard set by the SAEBRS-TRS TB scale in terms of sensitivity, a review of Study 1 instrumentation suggests alterations could be made to the SAEBRS-TN forms in the interest of enhancing diagnostic accuracy. As was previously described, SAEBRS-TN instructions called for teachers to select up to five of their students who most closely fit each definition of behavioral risk. By their nature, such instructions both (a) limit the possible number of students who could be considered via Gate 2 to five, while also (b) not requiring teachers to nominate any students as potentially at risk. Such instructions were intended to allow for the detection of students who were potentially at risk, while increasing the efficiency of the overall screening procedure by limiting the possible number of students for whom additional assessment would be required. Yet, the manner in which instructions were worded may have actually increased the potential for false negative decisions. A review of SAEBRS-TN descriptive statistics supported this potential, indicating that a rather large percentage of teachers nominated only 0–3 students for social behavior risk (61.0%), academic behavior risk (48.8%), and emotional behavior risk (82.9%).

Recognition of the ways in which SAEBRS-TN instrumentation might have influenced the tool's diagnostic accuracy suggested the evaluation of an alternative nomination procedure. Such an alternative approach might incorporate revised instructions intended to “cast a wider net” by requiring teachers to nominate a larger number of students and therefore evaluate a greater percentage of their students using the SAEBRS-TRS. The purpose of Study 2 was to replicate Study 1 procedures, while incorporating a revised SAEBRS-TN process that required teachers to nominate five or more students in their classrooms. It was hypothesized this alternative approach would yield increased sensitivity and thus serve as a more viable candidate as an initial gate within a SAEBRS-based multiple gating procedure.

3. Study 2

3.1. Method

3.1.1. Participants

Elementary ($n = 712$) and Middle ($n = 822$) school students from the southwestern United States participated in Study 2 (an additional 22 students opted out of participation; 98.59% response rate). Students were sampled from four schools across two districts, including two elementary schools (grades K–5) and two middle schools (grades 6–8). Participants also included 33 elementary teachers and 38 middle school teachers. Additional demographic information is provided in Table 1.

3.1.2. Measures, procedures, and data analyses

Study 2 measures, procedures, and data analyses were analogous to those employed as part of Study 1. No alterations were made to brief screening rating scales, parental consent processes, teacher trainings, or data collection procedures. However, in an effort to address limitations in Study 1, slight changes were made to the SAEBRS-based multiple gating procedure (as described above). An additional alteration to Study 1 procedures pertained to the approach to SAEBRS-TRS cut score selection and evaluation. Whereas Study 1 incorporated a combination of calibration and cross-validation procedures, Study 2 employed the cross-validation approach alone. This alternative approach was considered permissible in the presence of Study 1 findings, which permitted the evaluation of a single set of cut scores previously identified as suitable for use in universal screening. Specific cut scores evaluated as part of Study 2 include SB ≤ 12, AB ≤ 9, EB ≤ 17, and TB ≤ 36. The remaining data analysis steps were commensurate with those conducted in Study 1. Within Study 2, 15.2% of elementary and 12.2% middle school students were at risk for behavioral and emotional concerns per the BESS.

3.2. Results

3.2.1. Missing data

Data were reviewed to identify the extent and nature of data missingness across the SAEBRS-TRS, SAEBRS-TN, and BESS measures, prior to analyses. Overall, less than 1% of SAEBRS-TRS and BESS data (i.e., 0.30%) were missing from the dataset. Although results of Little's test suggested these data were not MCAR ($\chi^2 = 3041.26$, $p < .001$), such a small extent of missing data may be considered negligible and potentially ignorable (Kline, 2011). This would ultimately support the use of traditional missing data

Table 7
Study 2 reliability, validity, and diagnostic accuracy coefficients.

Scale	Cronbach's alpha	Correlation with BESS	Area under the curve (AUC)
Elementary school			
Social Behavior	.89	−.85**	.95** (.93–.96)
Academic Behavior	.92	−.88**	.96** (.93–.96)
Emotional Behavior	.82	−.75**	.89** (.86–.92)
Total Behavior	.94	−.94**	.98** (.97–.99)
Middle school			
Social Behavior	.88	−.83**	.92** (.89–.95)
Academic Behavior	.93	−.88**	.94** (.92–.97)
Emotional Behavior	.79	−.72**	.88** (.84–.91)
Total Behavior	.93	−.94**	.98** (.97–.99)

Note. SAEBRS-TRS = Social, Academic, and Emotional Behavior Risk Screener; BESS = Behavioral and Emotional Screening System.
** $p < .001$.

handling procedures, such as single imputation techniques. As such, similar to Study 1, missing data were handled via expectation–maximization to maintain the original sample of 1534 students.

An evaluation of SAEBRS-TN data revealed nomination data were missing for 256 student participants (16.69% of the sample), as many teachers did not complete the nomination survey package as it was a separate link from the BESS and SAEBRS-TRS survey. This suggested SAEBRS-TN data were missing at random (MAR). Similar to Study 1, listwise deletion was once again considered acceptable given previous research indicating use of this particular approach is appropriate when missingness is less than 30% (Janssen et al., 2010).

3.2.2. RQ #1

Table 7 contains the obtained reliability coefficients for Study 2. At the elementary level, all Cronbach's alpha values were found to exceed the .80 threshold for low-stakes decision making (range = .82–.94; Cortina, 1993; Nunnally, 1978). Middle school values were also consistent with these findings, with the exception of EB, which yielded a Cronbach's alpha (.79) that fell just below the threshold. Across both grade levels, the broad TB scale was found to yield the highest Cronbach's alpha value (elementary = .94 and middle = .93), whereas the EB subscale yielded the lowest (elementary = .82 and middle = .79).

3.2.3. RQ #2

Validity findings are included in Table 7. All resulting correlation coefficients were in the expected direction across both grade levels, statistically significant at the $p < .001$ level, and again met or exceeded the aforementioned universal screening-specific threshold for high validity coefficient (i.e., .69). Of the SAEBRS scales, the broad TB scale yielded the highest correlations with the BESS (elementary = −.94 and middle = −.94), whereas the EB subscale yielded the lowest (elementary = −.75 and middle = −.72).

3.2.4. RQ #3

Correlations between the SAEBRS subscales fell in the moderate to large ranges, with the correlation between SB and AB equaling .71, SB and EB equaling .63, and AB and EB equaling .61. Results of logistic regression analyses are presented in Table 8. At both the elementary and middle school levels, each subscale was found to be statistically significant predictor of student risk ($p < .001$) after accounting for all other subscales.

3.2.5. RQ #4

Similar to Study 1, the diagnostic accuracy of each SAEBRS scale (per the AUC statistic) was statistically significant across both grade levels, and in the moderate (.70–.90) or high (≥.90) range (see Table 7). Consistent with reliability and validity findings, the highest AUC values corresponded to the overall TB scale. The sole moderate AUC values corresponded to the EB subscale (i.e., elementary = .89 and middle = .88). At the elementary level, 95% CIs were entirely in the high range for SB, AB, and TB, while in the moderate to high range for EB. Middle school results were similar, but with SB and EB yielding CIs in the moderate and high ranges, while AB and TB CIs remained in the high range.

SAEBRS-TRS diagnostic accuracy was also supported via conditional probability statistics (see Table 9). AB sensitivity and specificity values fell in the optimal range (>.90) across both elementary and middle school. For SB, sensitivity was acceptable in elementary and middle school (>.80) and specificity was optimal across both grade levels. Similar to Study 1, the EB and TB scales yielded adequately performing cut scores. Within EB, a cut score of 17 yielded acceptable sensitivity at the elementary level and optimal sensitivity at the middle school level. Specificity for the EB scales fell in the acceptable range in elementary school and approximated the acceptable range in middle school. Within TB, a cut score of 36 resulted in optimal specificity across both grade levels. Sensitivity was in the optimal range for the elementary level and in the acceptable range for middle school. See Table 10 for a graphical depiction of selected cut scores' sensitivity and specificity with associated 95% CIs.

Table 8
Study 2 logistic regression analyses.

	Estimate	SE	z	p
Elementary school				
Intercept	12.60	1.56	8.07	<.001
SB	−0.42	0.07	−5.87	<.001
AB	−0.62	0.09	−6.70	<.001
EB	−0.25	0.07	−3.77	<.001
Middle school				
Intercept	13.55	1.59	8.55	<.001
SB	−0.43	0.06	−6.65	<.001
AB	−0.58	0.08	−7.22	<.001
EB	−0.32	0.07	−4.88	<.001

Consistent with results from Study 1, cross-validation cut scores demonstrated negative predictive values approaching 1.00 and comparatively lower positive predictive value (relative to negative predictive value), indicating that many students identified as at risk through SAEBRS scales were actually not truly at risk. Of the SAEBRS scales, TB continued to perform best, as positive predictive values exceeded those associated with any of the SB, AB, and EB scales. The positive predictive value of the EB scale was lower than the alternative scales.

Large positive likelihood ratios (>10) were noted for TB at both grade levels, moderate (5–10) for SB and AB at both grade levels, and small (2–5) for EB at both grade levels. In regards to negative likelihood ratios, large values (<0.1) were noted for SB and TB at the middle school level, moderate (0.1–0.2) for AB and EB at both grade levels and TB at the elementary level, and small (0.2–0.5) for SB at the elementary level.

Base rates were calculated for each SAEBRS-TRS scale to determine what percentage of students across the total sample would have been classified as at risk per the selected cut scores. Elementary level base rates were as follows: SB = 17.9%, AB = 20.8%, EB = 29.5%, and TB = 19.3%. Middle school level base rates were as follows: SB = 18.1%, AB = 20.3%, EB = 20.0%, and TB = 15.2%.

3.2.6. RQ #5

Conditional probability statistics associated with each of the three SAEBRS-TN forms (i.e., Social, Academic, and Emotional), as well as the Combined nomination approach, are provided in Table 9. Consistent with Study 1, the three SAEBRS-TN forms yielded optimal specificity (.86–.92), negative predictive values approaching 1.00 (.90–.95), and unacceptable levels of sensitivity (.42–.69). Positive predictive values were also less than adequate. Moderate positive likelihood ratios were found at both grade levels for the Social and Emotional scales, whereas small positive likelihood ratios were noted for AB at the Elementary level. All negative likelihood ratio values fell above the threshold for small. In contrast, to the individual SAEBRS-TN forms, the Combined nomination approach yielded borderline to acceptable sensitivity (.74–.89). This was in addition to acceptable specificity (.76–.76), less than adequate positive predictive values (.33–.35), negative predictive values approaching 1.00 (.95–.98), small positive likelihood ratios (3.03–3.77), and small to moderate negative likelihood ratios (.14–.34).

3.2.7. RQ #6

As in Study 1, the unacceptable levels of sensitivity afforded by the individual SAEBRS-TN forms suggested they were not suited for consideration within their respective multiple gating procedures. Yet, the performance of the Combined nomination approach suggested such consideration was once again appropriate. As indicated in Table 9, the Combined multiple gating procedure yielded borderline to acceptable sensitivity (.70–.81), optimal specificity (.96–.96), improved positive predictive values similar to those associated with the SAEBRS-TRS TB scale (.72–.73), negative predictive values approaching 1.00 (.95–.97), large positive likelihood ratios (15.78–18.87), and small to moderate negative likelihood ratios (.20–.31). To note, if the Combined

Table 9
Study 2 conditional probability statistics associated with Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) teacher rating scales, teacher nomination approaches, and multiple gating procedure.

SAEBRS scale	Cut	SE	SP	PPV	NPV	LR+	LR−
Elementary school							
Social	12	.86	.88	.56	.97	7.17	0.16
Academic	9	.93	.86	.55	.98	6.64	0.08
Emotional	17	.88	.72	.36	.97	3.14	0.17
Total	36	.97	.93	.71	.95	13.86	0.03
Teacher Nomination	Social	.50	.92	.51	.92	6.11	0.54
	Academic	.55	.86	.39	.92	3.87	0.52
	Emotional	.42	.92	.47	.90	5.27	0.63
	Combined	.74	.76	.33	.95	3.03	0.34
MGP	Combined	.70	.96	.72	.95	15.78	0.31
Middle school							
Social	12	.80	.90	.53	.97	7.90	0.23
Academic	9	.91	.85	.45	.99	6.07	0.11
Emotional	17	.90	.65	.26	.98	2.57	0.15
Total	36	.88	.92	.63	.98	11.00	0.13
Teacher Nomination	Social	.58	.90	.44	.94	5.61	0.47
	Academic	.69	.88	.44	.95	5.60	0.35
	Emotional	.55	.89	.42	.93	5.17	0.51
	Combined	.89	.76	.35	.98	3.77	0.14
MGP	Combined	.81	.96	.73	.97	18.87	0.20

Note. SE = sensitivity, SP = specificity, PPV = positive predictive value, NPV = negative predictive value, LR+ = positive likelihood ratio, LR− = negative likelihood ratio, and MGP = multiple gating procedure.

multiple gating procedure was used within the current sample, 13.8% of students would have been identified as at risk at both the elementary and middle school levels.

4. General discussion

The primary goals of this investigation were to (a) evaluate the technical adequacy of the SAEBRS-TRS across two settings, and (b) determine the defensibility of a streamlined universal screening system, which incorporated teacher nomination (Gate 1) and the SAEBRS-TRS (Gate 2) within a broader multiple gating procedure. Consistent with the results from Study 1, findings from Study 2 supported the overall internal consistency (RQ #1), concurrent validity (RQ #2), and diagnostic accuracy (RQ #4) of the SAEBRS-TRS within a different setting (i.e., southwestern United States). Findings also once again supported the unique contribution of each SAEBRS subscale relative to the prediction of student risk, suggesting each might play an individual role in evaluating student behavior and potential for behavioral and emotional concerns (RQ #3). Importantly, SB and AB cut scores identified via previous investigations (≤12 and ≤9, respectively; Kilgus et al., 2013, 2015) yielded acceptable to optimal performance in terms of sensitivity and specificity. Findings therefore speak to the robustness of these scores and their potential for use across multiple settings and grade levels in defining student risk. Similarly, the TB cut score of ≤36 performed adequately across both elementary and middle school samples within both studies, yielding preliminary evidence of the score's defensibility. In contrast, though the EB cut score of ≤17 performed well in both the Study 1 elementary and middle school samples and in the Study 2 elementary sample, its specificity fell below the acceptable range within the Study 2 middle school sample.

Such variation in performance of the EB cut score is not entirely unexpected, particularly for two reasons. First, variation in cut score performance has been documented in other areas of research, such as curriculum-based measurement of oral reading fluency (Kilgus, Methe, et al., 2014). Fortunately, such variability has not appeared to limit widespread implementation of such measures within multi-tiered systems of support initiatives, but rather justifies the need for schools to consider the most applicable cut scores on a local level. Second, differences in cut scores may be expected given the various developmental trajectories of behavioral risk, specifically with internalizing and emotional disorders, from elementary to middle school aged children (Masten et al., 2005). As children age it is reasonable to expect a greater degree and frequency of internalizing type behaviors (e.g., depression, stress, anxiety). Thus, it may be desirable to allow cut scores to vary and become more lenient across grade levels (i.e., elementary versus middle). Results supported this notion, as the EB cut score of 16 was found to yield acceptable performance (SE = .80 and SP = .79) within the Study 2 middle school sample. Taken together, though it is desirable to simplify the adoption of universal screening through the provision of similar cut scores across grade levels, the need to identify alternative cut scores based upon grade and developmental level should not preclude the adoption of a universal screening tool. Rather, such findings speak to the potential need for contextualization and further research in the interest of identifying optimal cut scores specific to each grade and developmental level.

A second purpose of the present investigation was to examine the usability of the SAEBRS-TRS via an evaluation of a multiple gating procedure. Results from Study 1 indicated an overall lack of sensitivity for the three SAEBRS-TN forms (RQ #5). This was likely a result of under-identification, with a large number of teachers nominating only a few students (e.g., 0–3) within their classroom per each nomination form. The SAEBRS-TN findings ultimately prevented further examination of the utility of the individual SAENRS-TN forms as part of broader multiple gating procedures. In contrast, the Combined nomination procedure yielded adequate performance indicative for its potential as a first gate, despite not reaching the optimal sensitivity levels that would be desired of a Gate 1 measure. Follow-up analyses confirmed the Combined multiple gating procedure was superior, given its improved conditional probability values (RQ #6). Results suggested there was room for improvement in the teacher nomination process, particularly as it pertained to the individual SAEBRS-TN forms and their corresponding construct-specific multiple gating procedures.

Thus, given (a) the potential variability in base rates of student risk (i.e., some schools may have significantly higher proportions of students exhibiting mental health symptomology), and (b) the observed low nomination rates (e.g., emotional), the

Table 10
Study 2 sensitivity (SE) and specificity (SP) values with corresponding bootstrapped 95% confidence intervals (CI-95) for selected SAEBRS-TRS cut scores.

SAEBRS-TRS scale	SE	SP	SE CI-95	SP CI-95
Elementary school				
Social Behavior	.86	.88	.80–.93	.85–.91
Academic Behavior	.93	.86	.88–.97	.83–.89
Emotional Behavior	.88	.72	.81–.94	.69–.76
Total Behavior	.97	.93	.94–1.00	.91–.95
Middle school				
Social Behavior	.80	.90	.71–.87	.88–.92
Academic Behavior	.91	.85	.85–.97	.82–.87
Emotional Behavior	.90	.65	.84–.95	.62–.68
Total Behavior	.88	.92	.81–.94	.91–.95

directions on the nomination form was changed from "up to five students" to "five or more students" in Study 2 to potentially improve the sensitivity of teacher nomination. Disappointingly, and consistent with the results from Study 1, the three SAEBRS-TN forms did not exhibit adequate sensitivity, once again precluding further evaluation within their proposed multiple gating procedures. In contrast, the Combined multiple gating procedure performed adequately across the elementary and middle school levels, with sensitivity falling in the borderline and acceptable ranges.

Taken together, results of both studies suggest that at this time, universal screening with the SAEBRS-TRS is more preferable and psychometrically defensible relative to the teacher nomination approaches and multiple gating procedures. Results from the present investigation were consistent with prior research that favors the defensibility of universal screening procedures over the potential improvements in usability via incorporation of teacher nomination (Dowdy, Doane, Eklund, & Dever, 2011; Lane, 2003). Yet, the current results were promising for the Combined multiple gating procedure, suggesting its use might be appropriate for use within applied settings. Future research is needed to both corroborate the current findings and determine whether modifications to the teacher nomination process might enhance diagnostic accuracy.

While the present change in wording did not result in improved sensitivity for the Combined multiple gating procedure, several possibilities exist to improve upon teacher nomination. For example, future studies may examine the efficacy of teacher training to enhance behavioral ratings (Kilgus, Kazmerski, Taylor, & von der Embse, in press). Research has supported the utility of systematic training to improve identification of and referral to services for mental health problems (e.g., Mental Health First Aid [MHFA]; Rothi, Leavey, & Best, 2007). Moreover, teacher training and specification of behavioral constructs (i.e., providing specific examples/non-examples via video vignettes) may improve ratings, especially given the difficulty in observing and identifying internalizing problems (Levitt & Merrell, 2009). For example, teachers may be provided explicit training in mental health symptomology and behavioral constructs at the beginning of the school year. Teachers may then be better informed of and primed to recognize symptoms as they interact with their students. These procedures may mitigate the challenges inherent within most behavioral rating scales, such as asking teachers to rate behaviors over the past months for which they may not have sufficient prior knowledge, and improve teachers' abilities to recognize risk via nomination methods.

Beyond teacher training, future studies might also consider alternative approaches to teacher nomination. This might include (a) requiring teachers to rank order their students with regard to potential risk for behavioral problems (in a manner consistent with the SSBD), or (b) requiring teachers to complete a brief single rating indicating the extent to which they feel each student meets each operational definition for behavioral risk. It is believed both of these procedures are inherent to the teacher nomination procedure examined herein. By selecting a student for nomination, a teacher is likely implying that the student would be rated highly in that area of risk (in an absolute fashion) and ranked highly relative to his or her peers (in a relative fashion). Yet, it might be advantageous to require teachers to explicitly provide such information. Either approach would lend themselves to calibration-based methodologies to ROC curve analysis, wherein it would be possible to identify either (a) the optimal number of students to be selected from the rank ordering, or (b) the cut score along the brief single rating scale that would differentiate those students who should pass through Gate 1 from those who should not.

4.1. Limitations

There are several limitations within the present investigation. First, the generalizability of the findings from both studies may be limited by the sample demographics. While the present investigation includes a relatively diverse sample across two settings, the results were limited by grade level (i.e., did not include high school students) and the time in which assessments were administered (i.e., behavioral and emotional risk may be more apparent later in the academic year). Longitudinal research, within and across academic years, will likely provide important data in the evaluation of decisional accuracy and predictive validity across time. Second, the teacher nomination procedure may have been influenced by ordering effects. Though the BESS and SAEBRS were randomly distributed, it was not possible to randomize the order in which teachers completed the nomination form and screening instruments (e.g., BESS and SAEBRS-TRS). Future research should consider alternate modes (e.g., paper copies) of assessment distribution to counterbalance the potential priming of teacher responses. Third, teachers may not have had adequate knowledge of the behavioral constructs to nominate students who were at risk with acceptable accuracy. The teacher nomination procedure (and rating scales) assumes opportunity to observe the behavioral construct; there are certain behaviors (e.g., withdrawal, worry) that may be setting specific (e.g., playground, lunchroom) and not typically observable by the classroom teacher. Fourth, given the limited sample of participating teachers, it was not possible to examine the nested structure of the data (i.e., students within classrooms within schools). Future research should employ larger samples of students and multiple raters (i.e., student, parent, and teacher ratings) to counteract said biases as well as accounting for the nested data structure. Such data is also necessary to examine measurement invariance and assess the equivalence of scores across subgroups.

Lastly, though the BESS has been used as a criterion outcome measure in several screening studies (e.g., Chafouleas et al., 2013; Kilgus et al., 2012), other more comprehensive assessment measures and procedures might represent more defensible indicators of student behavioral functioning. Such alternatives might include broadband behavior rating scales (e.g., BASC-2, TRF), or comprehensive multi-assessment diagnostic assessment procedures. Though the BESS might be considered a proxy of one of these measures (i.e., the BASC-2), the actual use of these measures would be preferred. It is therefore recommended that future researchers employ alternative and potentially more defensible outcome measures, allowing for a more rigorous evaluation of SAEBRS validity and diagnostic accuracy. Criteria of particular interest are those well aligned with the constructs assessed via SAEBRS subscales, including the SSIS, Academic Competence Evaluation Scales (DiPerna & Elliott, 2000), and the Devereux Student Strengths Assessment (DESSA; LeBuffe, Shapiro, & Naglieri, 2009). Such measures ought to also include criterion measures beyond those that are (a) completed by teachers, and (b) founded upon rating scale methodology. Such criteria might represent more appropriate outcomes against which to compare the teacher nomination approaches and multiple gating procedures, as teacher rating scales (e.g., the BESS) are likely to always be more associated with the SAEBRS-TRS as a result of mono-method and mono-informant biases.

4.2. Implications for practice

Prior research has supported SAEBRS-TRS psychometric defensibility, extrapolation inferences regarding the prediction of academic achievement and behavioral functioning, and decisional inferences regarding the reliable detection of students at risk for behavioral and emotional concerns (Kilgus et al., 2013, 2015; Kilgus, Kazmerski, et al., in press; Kilgus, Sims, et al., in press; von der Embse et al., in press). Results from the present investigation contribute to the robust extant evidence base for the applied use of the SAEBRS-TRS, through cross-validation with additional samples and settings, via diagnostic accuracy analyses, and with an initial examination of screening efficiency within a multiple gating approach. Taken together, the existing evidence represents increasing support for the applied use of the SAEBRS-TRS as a universal screening instrument within multi-tiered systems of support initiatives.

Specifically, research appears to support use of the SAEBRS-TRS to inform three types of decisions. First, educators could use the SAEBRS-TRS to determine which students are at risk for behavioral and emotional concerns and thus require Tier 2 or 3 services. Second, both the current logistic regression findings and prior validity research (Kilgus et al., 2013; Kilgus, Kazmerski, et al., in press; Kilgus, Sims, et al., in press) suggest the SAEBRS-TRS might inform educator decisions regarding the nature of each student's risk. Admittedly, universal screeners are commonly not expected to inform these latter decisions, being limited to informing the former type (Keller-Margulis, Shapiro, & Hintze, 2008). Yet, the documented psychometric defensibility (e.g., reliability, validity, diagnostic accuracy) of not only the broad TB scale, but also the three narrow subscales, suggests the SAEBRS-TRS might be used to identify the behavioral domain within which a student is struggling (i.e., social, academic, or emotional). These results could be used to guide intervention-related decisions, such as which behaviors should be targeted and what type of supports should be provided. For instance, if SAEBRS-TRS results suggest a student is struggling in the area of social behavior, educators might choose to engage in social skills training to support skill acquisition, as well as Check In/Check Out to reinforce the display of acquired skills. Similarly, documentation of struggles in the emotional behavior domain could support the psychoeducation of alternative coping and problem solving skills that might replace maladaptive internalizing behaviors.

The use of the SAEBRS-TRS informs decisions regarding (a) the presence of risk, and thus the need for intervention; and (b) the nature of risk, and therefore the type of intervention to be provided. A third type of decision the SAEBRS-TRS might inform corresponds to the prevalence of risk (i.e., base rates), and thus the level at which intervention should be provided. Kilgus and Eklund (2016) described how the SAEBRS-TRS could be used to identify base rates of risk across the various behavioral domains and at school- and class-wide levels. Documentation of increased base rates at the school level suggest the need for the implementation or modification of universal supports at Tier 1, such as the instruction and reinforcement of school-wide expectations (given social behavior concerns) or implementation of social–emotional learning programs (given emotional behavior concerns). Documentation of acceptable school-wide base rates but increased base rates within certain classrooms would support the need to modify those specific instructional environments, such as bolstering classroom management practices or implementation of class-wide interventions (e.g., Good Behavior Game). Lastly, documentation of acceptable base rates at the school- and class-wide levels would support the application of group- or individual-level Tier 2 interventions, such as social skills training or Check In/Check Out. Using screening base rate data to guide intervention (a) emphasizes the influence of the educational ecology on student behavior, stressing the need to manipulate environmental contingencies in support of behavioral change, and (b) supports a resource-conscious approach to service delivery, wherein costly Tier 2 interventions are only applied when less costly school- and class-wide interventions are documented as unnecessary.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Bruhn, A. L., Woods-Groves, S., & Huddle, S. (2014). A preliminary investigation of emotional and behavioral screening practices in K-12 schools. *Education and Treatment of Children*, *37*, 611–634.

Burke, M. D., Davis, J. L., Lee, Y. H., Hagan-Burke, S., Kwok, O. M., & Sugai, G. (2012). Universal screening for behavioral risk in elementary schools using SWPBS expectations. *Journal of Emotional and Behavioral Disorders*, *20*, 38–54.

Burns, M. K., Riley-Tillman, T. C., & VanDerHeyden, A. (2012). *RTI applications: Vol. 1. Academic and behavioral interventions.* New York: Guilford Press.

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

Carran, D. T., & Scott, K. G. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education*, *12*, 196–211.

Chafouleas, S. M., Riley-Tillman, T. C., & Sugai, G. (2007). *School-based behavioral assessment: Informing intervention and instruction.* New York: Guilford Press.

Chafouleas, S. M., Kilgus, S. P., Jaffery, R., Riley-Tillman, T. C., Welsh, M., & Christ, T. J. (2013). Direct Behavior Rating as a school-based behavior screener for elementary and middle grades. *Journal of School Psychology*, *51*, 367–385.

Conners, C. K. (1997). *Conners' Rating Scales – Revised: Long form.* North Tonawanda, NY: Multi-Heath Systems.

Cook, C. R. (2012). *The student externalizing behavior screener.* Unpublished document University of Washington.

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

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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

Daniels, B., Volpe, R. J., Fabiano, G. A., & Briesch, A. M. (2016). Classification accuracy and acceptability of the Integrated Screening and Intervention System Teacher Rating Form. *School Psychology Quarterly* (in press).

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

DiCenso, A., Guyatt, G., & Ciliska, D. (2014). *Evidence-based nursing: A guide to clinical practice*. Elsevier Health Sciences.

DiPerna, J. C., & Elliott, S. N. (2000). *Academic Competence Evaluation Scales*. San Antonio, TX: The Psychological Corporation.

Dowdy, E., Doane, K., Eklund, K., & Dever, B. V. (2011). A comparison of teacher nomination and screening to identify behavioral and emotional risk within a sample of underrepresented students. *Journal of Emotional and Behavioral Disorders*, *21*, 127–137.

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

Elias, M. J., & Haynes, N. M. (2008). Social competence, social support, and academic achievement in minority, low-income, urban elementary school children. *School Psychology Quarterly*, *23*, 474–495.

Elliott, S. N., & Gresham, F. M. (2008). *Social skills improvement system: Performance screening guide*. Minneapolis, MN: Pearson.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

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

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581–586.

Gresham, F. M., & Elliott, S. N. (2008). *Social skills improvement system: Rating scales*. Bloomington, MN: Pearson.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*, *34*, 372–386.

Janssen, K. J., Donders, A. R., Harrell, F. E., Vergouwe, Y., Chen, Q., Grobbee, D. E., ... Moons, K. G. M. (2010). Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*, *63*, 721–727.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, *36*, 582–600.

Jenkins, L. N., Demaray, M. K., Wren, N. S., Secord, S. M., Lyell, K. M., Magers, A. M., ... Tennant, J. (2014). A critical review of five commonly used social–emotional and behavioral screeners for elementary or secondary schools. *Contemporary School Psychology*, *18*, 241–254.

Johnson, A. J., Miller, F. G., Chafouleas, S. M., Welsh, M. E., Riley-Tillman, T. C., & Fabiano, G. (2016). Evaluating the technical adequacy of DBR-SIS in tri-annual behavioral screening: A multisite investigation. *Journal of School Psychology*, *54*, 39–57.

Kamphaus, R. W. (2012). Screening for behavioral and emotional risk: Constructs and practicalities. *School Psychology Forum*, *6*, 89–97.

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

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review*, *37*, 374–390.

Kilgus, S. P., & Eklund, K. (2016). Consideration of base rates within universal screening for behavioral and emotional risk: A novel procedural framework. *School Psychology Forum*, *10*, 120–130.

Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., & Welsh, M. E. (2012). Diagnostic accuracy of Direct Behavior Rating Single Item Scales as a screener of elementary school students. *School Psychology Quarterly*, *27*, 41–50.

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*, 210–226.

Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., & von der Embse, N. P. (2014). *Social, Academic, and Emotional Behavior Risk Screener (SAEBRS)*. Minneapolis, MN: Theodore J. Christ & Colleagues.

Kilgus, S. P., Riley-Tillman, T. C., Chafouleas, S. M., Christ, T. J., & Welsh, M. E. (2014). Direct behavior rating as a school-based behavior universal screener: Replication across sites. *Journal of School Psychology*, *52*, 63–82.

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology*, *52*, 377–405.

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.

Kilgus, S. P., Kazmerski, J. S., Taylor, C., & von der Embse, N. P. (2016). Use of Direct Behavior Rating (DBRs) to collect functional assessment data. *School Psychology Quarterly* (in press).

Kilgus, S. P., Sims, W. A., von der Embse, N. P., & Taylor, C. N. (2016). Psychometric defensibility of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) within an elementary sample. *Assessment for Effective Intervention* (in press).

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Kwon, K., Kim, E., & Sheridan, S. (2012). Behavioral competence and academic functioning among early elementary children with externalizing problems. *School Psychology Review*, *41*, 123–140.

Lane, K. L. (2003). Identifying young students at risk for antisocial behavior: The utility of “teachers as tests”. *Behavioral Disorders*, *28*, 360–369.

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J., Weisenbach, J. L., ... Phillips, A. (2009). A comparison of systematic screening tools for emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders*, *17*, 93–105.

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

LeBuffe, P. A., Shapiro, V. B., & Naglieri, J. A. (2009). *The Devereux Student Strengths Assessment*. Lewisville, NC: Kaplan.

Levitt, V. H., & Merrell, K. W. (2009). Linking assessment to intervention for internalizing problems of children and adolescents. *School Psychology Forum*, *3*, 13–26.

Masten, A. S., Roisman, G. I., Long, J. D., Burt, K. B., Obradovic, J., Riley, J. R., ... Tellegen, A. (2005). Developmental cascades: Linking academic achievement and externalizing and internalizing symptoms over 20 years. *Developmental Psychology*, *41*, 733–746.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283–298.

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Pennefather, J. T., & Smolkowski, K. (2015). Validation of the Elementary Social Behavior Assessment: A measure of student prosocial school behaviors. *Assessment for Effective Intervention*, *40*, 143–154.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children—Second Edition (BASC-2)*. Circle Pines, MN: AGS.

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*, *46*, 343–366.

Rothi, D., Leavey, G., & Best, R. (2007). On the front-line: Teachers as active observers of pupils' mental health. *Teaching and Teacher Education*, *24*, 1217–1231.

Rutter, C., & Gatsonis, C. (2001). A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Statistics in Medicine*, *20*, 2865–2884.

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

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristic curves. *The Canadian Journal of Psychiatry/La Revue Canadienne de Psychiatrie*, *52*, 121–128.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*, 522–532.

von der Embse, N. P., Pendergast, 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* (in press).

Walker, H. M., Severson, H. H., Nicholson, F., & Kehle, T. J. (1994). Replication of the Systematic Screening of Behavior Disorders (SSBD) procedure for the identification of at-risk children. *Journal of Emotional and Behavioral Disorders*, *2*, 66–77.

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