Title

Confirmation of Models for Interpretation and Use of the Social and Academic Behavior Risk Screener (SABRS)

Journal

School Psychology, 30(3)

ISSN

2578-4218

Authors

Kilgus, Stephen P
Sims, Wesley A
von der Embse, Nathaniel P
et al.

Publication Date

2015-09-01

DOI

10.1037/spq0000087

Peer reviewed

Confirmation of Models for Interpretation and Use of the Social and Academic Behavior Risk Screener (SABRS)

Article *in* School Psychology Quarterly · September 2014

DOI: 10.1037/spq0000087 · Source: PubMed

CITATIONS
25

READS
218

4 authors:

Stephen Kilgus
University of Wisconsin–Madison
69 PUBLICATIONS **650** CITATIONS
SEE PROFILE

Wesley A Sims
University of California, Riverside
6 PUBLICATIONS **72** CITATIONS
SEE PROFILE

Nathaniel Von der Embse
University of South Florida
52 PUBLICATIONS **588** CITATIONS
SEE PROFILE

T. Chris Riley-Tillman
University of Missouri
92 PUBLICATIONS **1,756** CITATIONS
SEE PROFILE

School Psychology Quarterly
2015, Vol. 30, No. 3, 335–352

1045-3830/15/$12.00 http://dx.doi.org/10.1037/spq0000087

Confirmation of Models for Interpretation and Use of the Social and Academic Behavior Risk Screener (SABRS)

Stephen P. Kilgus and Wesley A. Sims
University of Missouri

Nathaniel P. von der Embse
Temple University

T. Chris Riley-Tillman
University of Missouri

The purpose of this investigation was to evaluate the models for interpretation and use that serve as the foundation of an interpretation/use argument for the Social and Academic Behavior Risk Screener (SABRS). The SABRS was completed by 34 teachers with regard to 488 students in a Midwestern high school during the winter portion of the academic year. Confirmatory factor analysis supported interpretation of SABRS data, suggesting the fit of a bifactor model specifying 1 broad factor (General Behavior) and 2 narrow factors (Social Behavior [SB] and Academic Behavior [AB]). The interpretive model was further supported by analyses indicative of the internal consistency and interrater reliability of scores from each factor. In addition, latent profile analyses indicated the adequate fit of the proposed 4-profile SABRS model for use. When cross-referenced with SABRS cut scores identified via previous work, results revealed students could be categorized as (a) not at-risk on both SB and AB, (b) at-risk on SB but not on AB, (c) at-risk on AB but not on SB, or (d) at-risk on both SB and AB. Taken together, results contribute to growing evidence supporting the SABRS within universal screening. Limitations, implications for practice, and future directions for research are discussed herein.

Keywords: behavior assessment, social behavior, universal screening

Universal screening for behavior risk is defined as the systematic evaluation of a population to identify individuals displaying early symptoms of behavioral disorders (Jenkins, Hudson, & Johnson, 2007). The broader purpose of universal screening, be it for academic, health, or behavioral concerns, is to identify students who are likely to display problematic behavior if left unsupported (Kamphaus, 2012). The importance of screening has been reflected within legislation and by professional organizations, both of which have established screening as a prerequisite for prevention and early intervention (National Research Council and Institute of Medicine, 2009). Recognition of this importance has resulted in increased interest in the development and evaluation of universal screening instruments.

Multiple authors have recently offered comprehensive reviews of available instruments (Lane, Menzies, Oakes, & Kalberg, 2012; Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007). Severson et al. (2007) identified several categories into which screening instruments may be placed, including (a) multiple gating procedures, (b) teacher nomination of problem students followed by completion of broadband rating scales, and (c) universal teacher evaluation of all students on common behavioral criteria. Research has been particularly plentiful of late regarding this final cate-

This article was published Online First September 29, 2014.

Stephen P. Kilgus and Wesley A. Sims, Department of Educational, School, and Counseling Psychology, University of Missouri; Nathaniel P. von der Embse, Department of Psychological, Organizational, and Leadership Studies of Education, Temple University; T. Chris Riley-Tillman, Department of Educational, School, and Counseling Psychology, University of Missouri.

Correspondence concerning this article should be addressed to Stephen P. Kilgus, Department of Educational, School, and Counseling Psychology, 16 Hill Hall, University of Missouri, Columbia, MO 65211. E-mail: kilguss@missouri.edu

gory of screeners, with such instruments being defined by common characteristics. First, such screeners frequently employ rating scale methodology, through which teachers rate the frequency of multiple student behaviors using Likert scaling. Second, such screeners tend to be highly efficient, incorporating a small number of items (e.g., less than 30) that may be completed within a brief amount of time (e.g., less than 5 min). Such brevity is intended to support true universal screening, through which all students may be evaluated in an identical and timely fashion (Levitt, Saka, Romanelli, & Hoagwood, 2007). Third, such screeners typically correspond to a small number of latent variables considered highly relevant to social and academic success, including externalizing behavior, internalizing behavior, attentional problems, social competencies, and academic competencies (Kamphaus, 2012; Masten et al., 2005; Walker, Irvin, Noell, & Singer, 1992).

A review of the literature reveals many instruments that fit within the category of universal teacher evaluation screeners. Research is particularly plentiful regarding the *Behavioral and Emotional Screening System* (BESS: Kamphaus & Reynolds, 2007), as well as the *Student Risk Screening Scale* (SRSS; Drummond, 1994). Studies have also begun to support the psychometric defensibility of several novel instruments. One such instrument, which recently demonstrated strong promise for use in universal screening, is the *Social and Academic Behavior Risk Screener* (SABRS; Kilgus, Chafouleas, & Riley-Tillman, 2013).

Social and Academic Behavior Risk Screener (SABRS)

Kilgus et al. (2013) developed and initially validated the SABRS within a sample of 243 elementary school students in southeastern United States. Fifty-four teachers rated students across two measures, including the *Social Skills Improvement System* (SSIS) teacher rating scale and an initial version of the SABRS. Exploratory factor analyses supported the extraction of two factors and retention of 12 SABRS items. Six of these items corresponded to a *Social Behavior* (SB) factor, whereas the remaining six items corresponded to an *Academic Behavior* (AB) factor. SB items pertained to behaviors that influence a student's ability to maintain age appropriate relationships with peers and adults. These included maladaptive behaviors representative of externalizing problems (e.g., temper outbursts) and adaptive behaviors representative of social competencies (e.g., cooperation with peers). AB items pertained to behaviors that influence a student's ability to be prepared for, participate in, and benefit from academic instruction. These included maladaptive behaviors representative of attentional problems (e.g., distractedness) and adaptive behaviors representative of academic competencies (e.g., production of acceptable work). This interpretation of each factor was reinforced by reliability coefficients, which supported the internal consistency of each factor's scores (coefficient α = .90–.94), and validity coefficients, which supported each factor's concurrent criterion-related validity relative to the SSIS (mean of Pearson's r = .72). Findings also supported the interpretation of a broader *General Behavior* (GB) scale inclusive of both SB and AB items, with results indicative of the reliability (alpha coefficient = .93) and validity (mean of Pearson's r = .79) of GB scores.

Kilgus et al. (2013) also afforded evidence of SABRS diagnostic accuracy, with receiver operating characteristic (ROC) curve analyses yielding preliminary recommendations for SABRS cut scores. Results suggested that students should be considered at-risk for social behavior problems if their SB score was equal to or less than 12, as this score was associated with adequate levels of sensitivity (*SE*) and specificity (SP) relative to both the SSIS Social Skills scale (*SE* = .87 and SP = .83) and the SSIS Problem Behaviors scale (*SE* = .97 and SP = .84). Results also indicated that students should have been considered at-risk for academic behavior problems if their AB score was equal to or less than 11. Though multiple AB scores appeared appropriate, 11 was associated with adequate performance across multiple SSIS scales, including SSIS Social Skills (*SE* = .85 and SP = .76) and SSIS Academic Competence (*SE* = .89 and SP = .75). Finally, results suggested that students should have been considered at-risk for general behavior problems if their GB score is equal to or less than 24, with that particular score performing adequately relative to SSIS Social Skills (*SE* = .94 and SP = .81), SSIS Problem Behaviors (*SE* = .92 and

SP = .80), and SSIS Academic Competence (*SE* = .84 and SP = .73).

Results from the Kilgus et al. (2013) investigation yield support for continued SABRS research, suggesting the SABRS incorporates certain strengths of existing universal teacher evaluation screeners. Similar to the SRSS, the SABRS includes few items, which increases the measure's efficiency in screening (Kilgus et al., 2013). Similar to the BESS, SABRS content corresponds to multiple variables known to predict academic and social success (e.g., externalizing problems, social competencies; Masten et al., 2005; Walker et al., 1992), thus supporting the measure's contextual relevance. Together, these features suggest the SABRS might represent a unique and valuable contribution to the body of existing universal screeners and is therefore deserving of additional research. The scope and direction of this research is informed by a SABRS-specific *interpretation/use argument*, which is founded upon the initial findings from Kilgus et al. (2013).

Models for SABRS Interpretation and Use

Model for Interpretation

Kane (2001, 2013) described an interpretation/use argument as a network of inferences leading from test scores to test interpretation and use. The SABRS interpretation/use argument specifies a model for the interpretation of SABRS scores, as well as a separate model for the manner in which the SABRS is to be used in applied settings. The aforementioned exploratory factor analyses, as well as the resulting reliability and validity coefficients, inform the SABRS *model for interpretation*. The model specifies a three-factor structure, with two narrow factors (i.e., SB and AB) and one broad factor (i.e., GB). The model further indicates the narrow factors are positively related to each other, such that a student with high levels of social behavior is likely to also exhibit high levels of academic behavior.

When conceptualized within a latent variable framework, the model for interpretation may be represented as a bifactor structure, with each SABRS item corresponding to one of the two narrow factors, as well as a broad GB factor. Having gained their prominence in early work regarding intelligence theory, bifactor models have recently regained attention within education and psychology (DeMars, 2013). In accordance with the model for interpretation, a SABRS bifactor model would permit the broad GB factor to account for covariance between all SABRS items. It would further permit the narrow SB and AB factors to account for residual covariance within item clusters after controlling for GB. The bifactor model is conceptually similar yet distinct from a higher order factor model. Whereas the latter might specify SB and AB mediate the relationship between the SABRS items and GB, the bifactor specifies that the relationship between the items and GB is distinct from and orthogonal to SB and AB.

Model for Use

The previously described diagnostic accuracy analyses, as well as evidence of the performance of SB, AB, and GB cut scores, inform two SABRS *models for use*. The first model corresponds to the broad GB scale, specifying that the SABRS may be used to place each student into one of two categories: (a) those at-risk for general behavior problems and (b) those not at-risk for general behavior problems. The second model corresponds to the narrow SB and AB scales, stating that the screener may be employed in determining whether a student is at-risk for social behavior problems, academic behavior problems, or both. More specifically, the second model for SABRS use states that each individual student may be placed into one of four categories. See Figure 1 for a representation of each expressed within a 2 × 2 matrix. In Category 1, students are not at-risk on both SB (>12) and AB (>11). In Category 2, students are not at-risk on SB (>12) but are

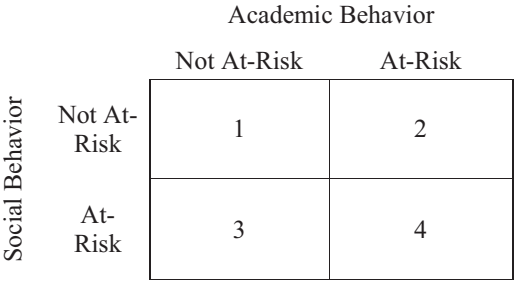

Figure 1. Model for applied use of the Social and Academic Behavior Risk Screener (SABRS).

at-risk on AB (≤11). This pattern is reversed for Category 3, where students are at-risk on SB (≤12) but not at-risk on AB (>11). Finally, in Category 4, students are at-risk on both SB (≤12) and AB (≤11).

Consistent with the aforementioned model for interpretation, the models for use may be conceptualized within a latent variable framework, wherein the variable of interest would represent a latent categorical variable indicative of student category membership (Pastor, Barron, Miller, & Davis, 2007). Through such a framework, categories within both models are considered categorical profiles of the latent variable. Each profile includes students displaying similar SABRS item scores, such that items within each profile are conditionally independent because of limited intraprofile variation. Common membership within a latent profile is assumed to account for similar performance across students, as the profiles are thought to "cause" item scores and thereby determine risk status (Pastor et al., 2007). In other words, the presence or absence of risk on either SABRS scale is assumed to influence the scores observed on each constituent item and thus determine whether the student will be considered at-risk when his or her observed total scale score is compared with total scale cut scores.

Summary

Taken together, the models comprising the SABRS interpretation/use argument provide guidance regarding how the SABRS should be applied within school settings. They also provide a framework through which additional psychometric research should be planned and conducted (Kane, 2001). Future research regarding the model for interpretation should expand the evidence supporting the construct validity of SABRS score-based inferences (Kane, 2013). Specifically, research should verify the proposed SABRS bifactor model via more advanced statistical methods, such as confirmatory factor analysis (CFA). Research should also corroborate existing reliability and validity evidence while examining additional reliability and validity types. It would be of interest to examine the interrater reliability of SABRS scores; that is, the extent to which SABRS data are consistent across teacher raters. Findings would have direct implications for SABRS procedures in school settings. High interrater reliability would indicate the potential interchangeability of SABRS raters and might support flexibility in rater selection. Low interrater reliability would call into question the defensibility of SABRS ratings and suggest the need for multiple raters of each student in the interest of decision verification.

Future research regarding the models for use should yield evidence of the consequential validity of SABRS score utilization (Kane, 2001). At this early stage of SABRS-related inquiry, such research should look to confirm the performance of aforementioned SABRS cut scores in differentiating between at-risk and not at-risk students. This might be accomplished via latent profile analysis (LPA), through which it would be possible to evaluate the extent to which the proposed latent models fit SABRS data. If LPA findings were to support the proposed models, analyses would reveal profiles that could be differentiated in terms of average performance relative to SABRS cut scores (e.g., von der Embse, Mata, Segool, & Scott, 2014). For instance, if a LPA were to confirm the aforementioned two-category model, then results would suggest adequate fit of a two-latent profile model, with the average GB score falling below the previously recommended cut score (i.e., 24) in Profile 1 and above this cut score in Profile 2.

Purpose of the Study

The broader purpose of the investigation was to collect evidence described in the two preceding paragraphs, as such information directly pertains to the defensibility of the SABRS interpretation/use argument. An additional purpose was to expand the SABRS literature to an alternative grade level (high school) and setting (Midwest). Three research questions related to the model for SABRS interpretation were of interest. First, to what degree is the model for interpretation supported within a novel sample at the high school level, as evaluated via CFA? Analytic procedures mirrored those employed by DiStefano, Greer, and Kamphaus (2013). Specifically, analyses compared a bifactor structure, which closely represented the proposed interpretive model, to conceptually similar but more parsimonious factor structures. These included a unidimensional factor model, which specified the broad GB factor but not the

narrow SB and AB factors, and a correlated factor model, which specified the narrow SB and AB factors but not the broad GB factor. It was hypothesized that the bifactor model would provide the best fit to the data. Second, to what extent are SABRS data reliable across raters within each of the SABRS factors? This was evaluated in via two approaches. The first approach employed correlational statistics in examining the reliability of continuous SABRS scores. The second approach was founded upon agreement statistics in evaluating the reliability of dichotomous SABRS scores. Dichotomous scores were calculated through the use of previously identified SABRS cut scores, with at-risk students receiving a score of 1 and not at-risk students a score of 0. Findings corresponded to interrater reliability of decisions regarding student risk status. It was hypothesized that data would be moderately reliable, with a degree of unreliability reflecting true differences in student behavior across settings and in relation to different raters. Third, to what extent are SABRS factors internally consistent? In accordance with the results of Kilgus et al. (2013), it was hypothesized each scale would yield high internal consistency.

An additional question pertained to the models for SABRS use. Specifically, to what extent did SABRS data fit the previously described models for use, as evaluated via LPA? It was hypothesized that findings would support the two- and four-profile models. It was further anticipated that the four-profile model would yield superior fit, thus supporting a more complex characterization of student behavior across multiple behavioral domains.

Method

Participants and Setting

All study procedures were conducted within a single high school located in a rural Midwestern school district (Grades 9–12). In total, 34 teachers rated the behavior of 488 students using the SABRS. Teachers completing SABRS forms included 23 females and 11 males, all of whom were White, non-Hispanic. Reported years of teaching experience ranged from 0 to 20 or more and level of professional training reported ranged from bachelor's to master's degree. Teachers instructed in a variety of areas, with 8 in Communications Arts (24%; Foreign Language, English/Language Arts), 4 Math (12%), 4 Practical Arts (12%; e.g., Agriculture, Shop, Mechanic, Drafting), 4 Science (12%), 4 Social Studies (12%), 3 Physical Education or Health (9%), 3 Special Education (9%), 2 Fine Arts (6%), and 2 Business (6%). Of the 488 students, 291 were male and 197 were female. Fewer than 1% of students identified as English language learners and 22.5% of students qualified for free or reduced lunch. With regard to race/ethnicity, 94.5% of students identified as White, non-Hispanic, 2.3% as African American, 1.2% as Asian American, and 1.2% as White, Hispanic.

A review of the literature suggested the current sample size was sufficient to support the CFA plan (MacCallum, Widaman, Zhang, & Hong, 1999). With regard to reliability analyses, the current sample was similar in size to recent studies examining internal consistency and interrater reliability (e.g., King, Reschly, & Appleton, 2012). Finally, though findings are limited and heuristics are lacking, research has yielded some recommendations regarding sample sizes necessary to support LPA. Notably, among other considerations, smaller sample sizes are acceptable when the number of profiles is small, the size of each profile is moderate to large, and the analysis is to include a limited number of manifest variables (Samuelsen & Raczynski, 2013). Recent investigations employing LPA, examining behavior rating scale data, and considering profiles similar in number and size to those expected within the current study have included sample sizes similar to those presented herein (e.g., Herman, Ostrander, Walkup, Silva, & March, 2007).

Measure

One or more teachers rated each student using the SABRS (Kilgus et al., 2013). Teachers rated items by indicating the frequency with which the student in question displayed the described behaviors during the past month. Ratings were completed using a 4-point Likert scale, with 0 = *Never*, 1 = *Sometimes*, 2 = *Often*, and 3 = *Almost Always*. Subsequent to completion of all ratings, scores on negatively worded items (e.g., 'Arguing') were reverse scored, such that ratings of '0' were transformed to '3' and ratings of '1' were trans-

formed to '2' (and vice versa). Total summed scores were then derived for all three scales. SB and AB total scores ranged between 0 and 18, whereas GB scores ranged between 0 and 36. Across each scale, higher scores were indicative of more adaptive functioning and fewer problem behaviors.

Procedures

Before the completion of SABRS ratings, administrators and student support staff provided teachers with a brief overview of screening procedures (approximately 10 minutes in duration). Teachers then completed the SABRS for each of the students enrolled in their various classes (approximately 2–3 min per student). All data collection took place during the fall portion of the 2012–2013 school year (i.e., November). Teachers were permitted to complete their ratings via an online survey system at any point during a 24-day period. Student support staff coordinated with teachers to ensure that each student was rated by at least two teachers. In some instances ($n = 48$), coordination was not possible, resulting in some students only being rated by one teacher. A review of collected data indicated that one teacher rating was available for 100% of students, 2 ratings for 90.16%, 3 for 72.13%, 4 for 43.03%, 5 for 18.65%, 6 for 3.48%, and 7 for 0.41%. The number of students rated by each teacher ranged between 6 and 72 ($M = 47.06$, $SD = 17.25$). The school staff conducted the procedures described above as part of normal educational practices. Following the school's use of the current data for their own universal screening purposes, all data were provided to the researchers in de-identified format in accordance with Institutional Review Board (IRB)-approved procedures.

Data Analysis Plan

Missing data and data organization. The manner in which data were organized, as well as the approach to missing data handling, was dependent upon the analysis of interest. These discrepancies were primarily attributable to differences in the data requirements for each set of analyses. Specifically, whereas interrater reliability analyses required ratings from two or more teachers, CFA and LPA required information from a single teacher. See below for information regarding the approach employed in addressing each of the research questions.

Interrater reliability. Before interrater reliability analyses, teacher raters were randomly assigned an order within each student profile. For example, if four teachers rated Student 22, the teachers were randomly assigned the roles of Teacher 1, Teacher 2, Teacher 3, or Teacher 4. Next, the first three teachers in this order were selected for consideration in interrater reliability analyses. Although more than three teachers rated some students, it was of interest to limit the extent of missing data. By limiting analyses to the consideration of only three teachers, only 27.87% of students had some missing data across one or two teachers (all students were rated by at least one teacher). This was considered to be preferable relative to the consideration of four teacher ratings, which would have resulted in an inflated missing data rate of 56.97%. The use of only two teacher ratings for each student was also considered, as this would have resulted in a missing data rate of only 9.84%. Yet, it was ultimately determined that the inclusion of additional teacher ratings was desirable, given consideration of more data and calculation of additional agreement statistics was likely to yield more generalizable and defensible conclusions regarding interrater reliability.

Missing data were then imputed using multiple imputation, resulting in complete data for all 488 students across three teachers. Twenty imputed datasets were generated using an imputation model inclusive of all 12 SABRS items. Analyses were then conducted within each of the 20 datasets and results were pooled using Rubin's rules (Rubin, 1987). The use of multiple imputation was supported by descriptive analyses, which were indicative of the approximately normal distribution of data within each item, as well as by the assumption that the current missing data were missing at random (MAR). The MAR mechanism is tenable when missing data on X are related to one or more measured variables but not to the underlying values of the X (Enders, 2010). A review of the data indicated that when data were not available for a student from more than one teacher, it was primarily a result of either difficulties in the coordination of ratings or concerns regarding the expenditure of teacher time and resources. In contrast, it was not assumed that the absence

of teacher ratings was related to student behavior, thereby suggesting item missingness would not be predicted by SABRS item values.

CFA, LPA, and Cronbach's alpha. A fundamental assumption of CFA, LPA, and Cronbach's alpha is the independence of observations. For analyses to operate in accordance with this assumption, it was necessary to first remove data dependency in the form of multiple teacher ratings per each student. This was accomplished via the random selection of one teacher's ratings for each student. For the hypothetical Student 22 noted in the previous subsection, this would be achieved via the random selection of ratings from only one of the four teachers who completed the SABRS for the student. As a result, each of the 488 students was only represented by a single teacher' ratings within the final dataset, thus removing data dependence and supporting further analysis. As no missing data were present in this dataset, no missing data handling techniques were required.

After random selection, the new dataset was reviewed to ensure the selection process resulted in fair distribution of teacher ratings, with no single teacher yielding a disproportionately low or high number of student ratings. This review suggested random selection yielded ratings from all 34 teachers. The number of student ratings from a single teacher ranged from 3 (0.61%) to 28 (5.74%), with a mean of 14.35 (2.95%) and standard deviation of 5.95. Furthermore, the distribution of the number of student ratings per teacher was found to be approximately normal (Skewness = −0.05, Kurtosis = 0.25), indicating the ratings represented in the new random selection dataset were representative of those in the original dataset (which was also approximately normally distributed).

Interrater reliability. Before the evaluation of SABRS interrater reliability, total scale scores were calculated for each student within the SB, AB, and GB factors. These total scale scores then served as the basis for each of the two methods by which interrater reliability was evaluated for continuous SABRS scores. The first of these was through the calculation of intraclass correlation (ICC) coefficients. Specifically, the single measure ICC was considered, providing an index of the reliability of SABRS total scores for a typical single rater. The single measure statistic was preferred over the average measure statistic in recognition of the manner in which the SABRS is likely to be applied, with each student being rated by a single teacher rather than multiple teachers. ICC coefficients were calculated via the one-way random effects method, through which raters are considered randomly selected from a population of raters. ICCs range between .00 and 1.00, with values below .40 representing poor reliability, .40 to .59 fair reliability, .60 to .74 good reliability, and .75 to 1.00 high reliability (Cicchetti, 1994). Interrater reliability was also evaluated via the calculation of Pearson product–moment correlation coefficients. Three Pearson's *r* coefficients were calculated within each SABRS scale, allowing for comparison of scores between Teachers 1 and 2, Teachers 1 and 3, and Teachers 2 and 3. Means and ranges of these coefficients were then reported.

Next, dichotomous scale scores were calculated to support evaluation of interrater reliability for SABRS risk scores. For SB, a student was considered at risk (= 1) if their continuous scale score was less than or equal to 12 and not at risk (= 0) if their scale score was greater than 12. For AB, a student was considered at risk (= 1) if their continuous scale score was less than or equal to 11 and not at risk (= 0) if their scale score was greater than 11. These dichotomous scale scores then served as the basis for each of the two methods by which interrater reliability was evaluated for SABRS risk scores. The first of these was via a percent agreement statistic, which was calculated as the number of students identified by both teachers as at risk (A) plus number of students identified by both teachers as not at risk (B), divided by the total number of students (C; [A + B]/C). The second method by which dichotomous score interrater reliability was evaluated was via Cohen's kappa (κ), which is defined as proportion of agreement between two measures corrected for chance. In accordance with parameters outlined by Forstmeier and Maercker (2007) and Lane et al. (2009), κ values less than .20 were considered as poor, .21 to .40 as fair, .41 to .60 as moderate, and ≥.61 as good. Similar to Pearson's *r* results, three agreement coefficients were calculated within each SABRS scale (Teacher 1 vs. 2, Teacher 1 vs. 3, and Teacher 2 vs. 3), with coefficients means and ranges then reported.

Internal consistency. The internal consistency of scores within each of the three pre-

sumed SABRS scales (i.e., SB, AB, and GB) was evaluated via the calculation of Cronbach's alpha coefficients. Although recommendations vary, it has been suggested that alpha coefficients of at least .80 are needed to support low stakes decisions, whereas values of at least .90 are needed to support high stakes decisions (Cortina, 1993; Nunnally, 1978).

Confirmatory factor analysis. The appropriateness of the previously described SABRS model for interpretation was evaluated via CFA. In a manner consistent with analytic procedures employed by DiStefano et al. (2013), multiple factor structures were specified in accordance with the general interpretive model. These factor structures were then compared to determine which fitted the data best. See Figure 2 for an overview of the three structures evaluated as part of this investigation. Model A represented a unidimensional model, through which each SABRS item loaded on a single broad factor representative of GB. Model B represented a correlated factors model, which specified that each SABRS item loaded on one of two narrow covarying factors (i.e., SB and AB). Model A and B represented parsimonious factor structures, through which either the hypothesized narrow or broad factors were not specified. Estimation of these models permitted evaluation of whether a more complex multifactor model was necessary to appropriately model the SABRS data. Model C represented this more complex structure, specifying both narrow and broad SABRS factors. Specifically, each item was modeled as corresponding to both one of the two narrow factors and the broad factor. This particular bifactor structure specified that the covariance among items could be explained by (a) a broad factor, as well as (b) a group of narrow factors that accounted for covariance among items beyond that which was explained by the broad factor.

All CFAs were conducted using Mplus Version 7.1. Factors were extracted within each of

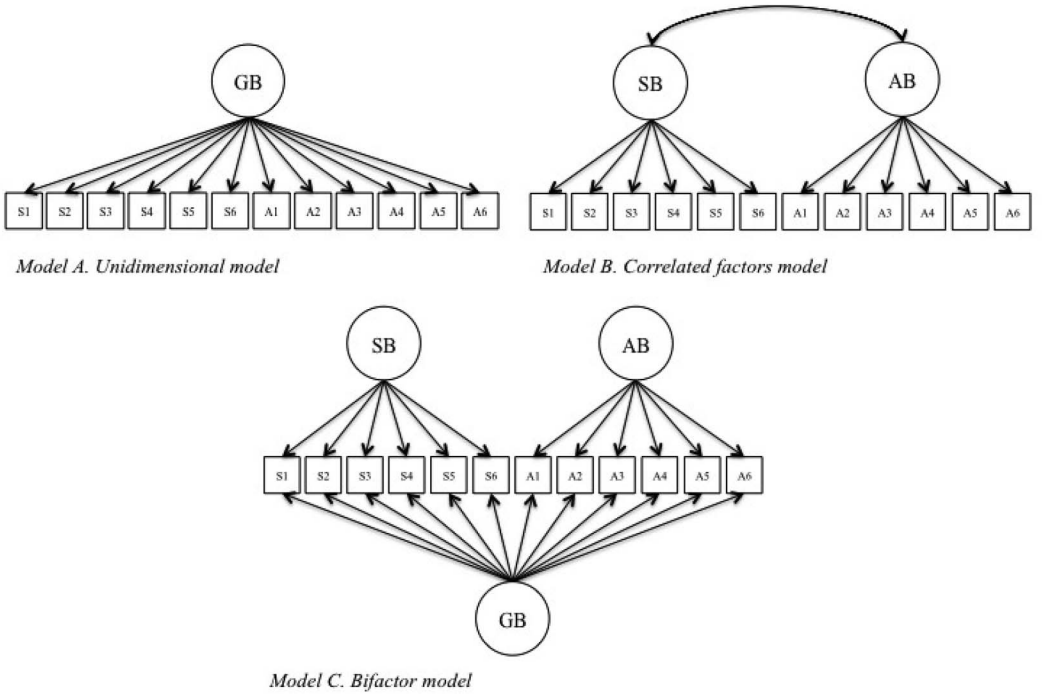

Figure 2. Factor models of the Social and Academic Behavior Risk Screener (SABRS) tested via confirmatory factor analysis. Item residuals deleted in the interest of visual simplicity. SB = Social Behavior, AB = Academic Behavior, and GB = General Behavior. Item abbreviations are indicative of each item's corresponding factor (S = Social and A = Academic) and item number.

models using maximum likelihood (ML) estimation, which assumes multivariate normality. A separate set of analyses was conducted using mean- and variance-adjusted weighted least squares (WLSMV) estimation, which is robust to violations of normality in the calculation of chi-square statistics and standard errors (Kline, 2011). ML and WLSMV findings were then compared through a sensitivity analysis, permitting an evaluation of the influence of normality violations on results. Results were equivalent across the two sets of analyses, thus supporting consideration of SABRS data as approximately normal and allowing interpretation of ML findings.

The fit of each factor model was evaluated via five fit statistics, including the chi-square goodness-of-fit test, Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), Comparative Fit Index (CFI; Bentler, 1990), Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980), and Standardized Root Mean Square Residual (SRMR; Bentler, 1995). Per recommendations from Hu and Bentler (1999), model fit was suggested in the presence of a nonstatistically significant chi-square test, CFI and TLI $\geq$ .95, RMSEA $\leq$ .06, and SRMR $\leq$ .08. Particular credence was given to the interpretation of the latter four statistics, as the chi-square test is generally considered to be an exceedingly stringent indicator of model fit, particularly with large sample sizes (Bentler & Bonnet, 1980).

Latent profile analysis. LPA was used in evaluating the fit of the previous described two- and four-profile models for SABRS use. Latent profile analysis is a specific case of finite mixture modeling that allows for classification of individuals into homogenous groups based on scores across multiple continuous variables (Lazarsfeld & Henry, 1968). Generally, the model selected as the best representation of the data has (a) the lowest Bayesian Information Criterion (BIC; Schwarz, 1978); (b) lowest Adjusted BIC, and Akaike's Information Criterion (AIC; Akaike, 1987); (c) highest entropy value (Hix-Small et al., 2004); and (d) utility in practice (Nagin, 2005). Entropy values range from 0 to 1, with higher values indicating a higher degree of certainty that an individual case is assigned within the correct profile. Researchers generally use one of two tests in evaluating relative model fit, including a Lo Mendell Rubin Likelihood Ratio Test (LMR LRT; Beyers & Seiffge-Krenke, 2007) or a bootstrapped LRT (BLRT; McLachlan & Peel, 2000). Simulation studies have produced equivocal results; Tofighi and Enders (2007) suggested that a combination of AIC, BIC, and LMR LRT to be used when evaluating model fit. However, recent simulation work by Nylund and colleagues (2007) found that BLRT consistently performed the best in reliably evaluating model fit and was therefore used within the present study. All LPAs were conducted using Mplus Version 7.1 (Muthén & Muthén, 2013).

Results

Item and scale level descriptive statistics were reviewed before conducting CFA, LPA, and reliability analyses (see Table 1). Items were found to be relatively uniform in terms of mean and standard deviation. Of note, scale score means fell above previously specified SABRS cut scores across all three scales, suggesting the average student exhibited behavior that would be classified as not at-risk.

Reliability

Internal consistency. Cronbach's alpha coefficients were indicative of the high internal consistency reliability of all three SABRS scale, with each value exceeding the minimal acceptable range of .70 to .80 (Cortina, 1993). Spe-

Table 1
Descriptive Statistics of Individual Items, as Well as Social Behavior, Academic Behavior, and General Behavior Scales

Item/scale	*M*	*SD*
S1	2.64	0.67
S2	2.50	0.72
S3	2.78	0.57
S4	2.61	0.70
S5	2.60	0.66
S6	2.52	0.78
A1	2.30	0.83
A2	2.32	0.83
A3	2.38	0.78
A4	2.50	0.82
A5	2.21	0.89
A6	2.31	0.85
Academic behavior	14.02	4.31
Social behavior	15.66	3.33
General behavior	29.67	6.94

cifically, Cronbach's alpha was equal to .89 for SB, .93 for AB, and .93 for GB.

Interrater reliability. ICCs were calculated within each of the three SABRS scales, permitting an evaluation of the reliability of continuous scores across three teacher raters. ICC coefficients were equal to .41 for SB (95% confidence interval [CI] = .37–.47), .46 for AB (95% CI = .41–.51), and .48 for GB (95% CI = .42–.53). All values fell in the "fair" range, suggesting a minimally acceptable level of interrater reliability. This finding was corroborated through a review of resulting Pearson's *r* coefficients. Correlations between teacher raters ranged between .35–.49 for SB (M = .41), .44 to .50 for AB (M = .47), and .45 to .51 for GB (M = .48). All *r* coefficients were statistically significant ($p < .05$).

Additional statistics were calculated to evaluate the reliability of dichotomous SABRS scores, or the extent to which teachers agreed regarding student risk status. The mean percent agreement across teachers was equal to 74.60% for AB (Range = 71.76–76.93%) and 81.61% for SB (Range = 81.25–82.03%). This indicated that teachers similarly classified approximately 3 out of every 4 students on AB and 4 of every 5 students on SB. Mean κ coefficients were indicative of fair agreement between teachers, with the mean κ value equal to .33 for AB (Range = .27–.37) and .31 for SB (Range = .25–.34).

Confirmatory Factor Analysis

Descriptive statistics were reviewed before conducting all CFAs in an evaluation of item normality, which was an assumption of the ML estimation method. Univariate normality was defined through cutoff values of ± 2.0 for skewness and ± 7.0 for kurtosis (Curran, West, & Finch, 1996). Although one of the 12 items was found to be non-normal (i.e., Temper outbursts), results of a sensitivity analysis supported the use of ML (see the Data Analysis section above for additional information regarding this analysis). Next, each of the models depicted in Figure 2 were examined and compared. See Table 2 for a summary of resulting model fit statistics. All five statistics were unanimously indicative of the ill-fitting nature of the unidimensional and correlated factor models. Greater support was found for the bifactor model (referred to as 'Bifactor 1' in Table 2), with observed SRMR, CFI, and TLI values meeting their corresponding cutoff values and RMSEA approaching its cutoff value.

Modification indices were therefore reviewed for the bifactor model in determining which parameters should be estimated toward the improvement of model fit. Findings supported the estimation of the covariance between two sets of item residuals, including (a) 'Cooperation with Peers' and 'Polite and socially appropriate responses toward others,' and (b) 'Interest in academic topics' and 'Academic engagement.' Each model revision was considered appropriate given the conceptual similarity between items and that items within each set corresponded to the same factor. Both adjustments were therefore made and a follow-up CFA was conducted. The resulting model ('Bifactor 2' in Table 2) provided good fit to the data, yielding fit statistic values that met their cutoff values for RMSEA, CFI, TLI, and SRMR.

See Table 3 for a summary of pattern coefficients associated with the adjusted bifactor model. Pattern coefficients may be interpreted in a manner consistent with beta weights resulting from a multiple regression, indicating the relationship between each item and factor after controlling for all

Table 2
Model Fit Statistics Associated With Various Factor Models

Statistic	Unidimensional	Correlated factors	Bifactor 1	Bifactor 2
χ^2	1299.497*	527.316*	203.432*	129.528*
RMSEA (90% CI)	.217	.135 (.074, .097)	.089 (.077, .101)	.068 (.055, .081)
SRMR	.119	.085	.044	.050
CFI	.728	.897	.965	.980
TLI	.668	.871	.945	.968

Note. χ^2 = Chi-square goodness-of-fit test; TLI = Tucker-Lewis Index (Tucker & Lewis, 1973); CFI = Comparative Fit Index (Bentler, 1990); RMSEA = Root Mean Square Error of Approximation (Steiger & Lind, 1980); and SRMR = Standardized Root Mean Square Residual (Bentler, 1995).
* Statistically significant at the $p < .001$ level.

Table 3
Pattern Coefficients Resulting From an Adjusted Bifactor Model, Indicative of the Relationship Between Each Item and Factor After Controlling for All Other Factors

Item	Social behavior	Academic behavior	General behavior
S1	.737		.589
S2	−.250		**−.593**
S3	.498		**.596**
S4	.255		**.801**
S5	−.347		**−.648**
S6	.134		**.804**
A1		.718	−.525
A2		.712	−.564
A3		.695	−.563
A4		−.270	**.710**
A5		−.280	**.814**
A6		.667	−.567

Note. Bolded values correspond to items yielding pattern coefficients that were higher for the broad General Behavior factor than the narrow Social Behavior and Academic Behavior factors.

other factors. Results suggested that the majority of SB items (i.e., 66.67%) were more strongly associated with the broad GB factor than their corresponding narrow factor. The opposite was true of AB items, the majority of which were better indicators of their narrow factor (i.e., 66.67%).

Latent Profile Analysis

LPA was used to evaluate the performance of the hypothesized two-profile and four-profile models for SABRS use. An additional three-profile model was also specified to examine the incremental fit of each of the hypothesized models. Its specification was necessary given that evaluations of relative model fit require competing LPA models to differ by only one profile. The authors evaluated goodness of fit indices (e.g., BIC, Entropy) and BLRTs to determine the optimal number of profiles. In the absence of a "gold standard," and consistent with similar methodology when determining model fit (e.g., Structural Equation Modeling), the authors identified the optimal level of profiles considering fit indices, theory, profile size, and uniqueness of profiles (Nagin, 2005; Pastor et al., 2007). Although potentially offering superior fit, a five-profile model (a) was not consistent with the hypothesized 2 × 2 risk model and (b) resulted in a profile with limited interpretability and usability (i.e., less than 5 students). Thus, the five-profile solution was ultimately rejected because of incongruence with theory, limited profile uniqueness, and small size of the fifth profile.

Table 4 contains the BLRT, BIC, Adjusted BIC, and Entropy values. In accordance with hypotheses, a four-profile model best represented the data, as it had the lowest BIC, Adjusted BIC, and AIC values, as well as a high entropy value. The four-profile model also yielded a statistically significant BLRT value, indicating its superior fit relative to the three-profile model (which was also superior to the two-profile model). Table 5 contains the average latent class probabilities for most likely latent class membership by latent class. A review of diagonal values indicated that on average, students had a high posterior probability of being assigned to their respective group. Predominantly low off-diagonal values further indicated that profiles were relatively distinct from one another.

Figure 3 represents the final four-profile model, depicting the mean of item scores across students falling within each profile. The first

Table 4
Fit Indices and Profile Prevalence (%) of the Latent Profile Analyses (n = 488)

						Proportion of students in the most likely class			
Solution	BIC	A-BIC	AIC	Entropy	BLRT	1	2	3	4
Two-profile	10978.18	10860.75	10823.14	.956	.00	.71	.29		
Three-profile	10032.11	9873.41	9822.59	.957	.00	.60	.30	.10	
Four-profile	9634.97	9435.01	9370.98	.966	.00	.56	.10	.26	.09

Note. The values in the BLRT column are the *p* values associated with BLRT in comparing fit between models. A-BIC = Sample-Size Adjusted BIC. The numerical numbered columns represent the percentages of students in the most likely class.

Table 5
Average Latent Class Probabilities for Most Likely Latent Class Membership (Row) by Latent Class (Column)

Class	1	2	3	4
1	.993	.002	.005	.000
2	.016	.951	.028	.005
3	.018	.006	.974	.002
4	.000	.007	.011	.982

Note. Columns refer to the latent class and rows refer to the most likely profile membership.

profile (n = 272; 55.7%) was associated with high scores across all AB and SB items (suggesting more appropriate behavior across both domains). A second profile (n = 48; 9.7%) evidenced high SB item scores and low AB item scores. A third profile (n = 125; 25.6%) was associated with high AB item scores and low SB item scores. Finally, a fourth profile (n = 43; 8.7%) evidenced low scores across all AB and SB items.

Next, mean item scores were summed within each scale, yielding two overall mean scores within each profile (i.e., one for SB and one for AB). Each mean factor score, which is depicted in Figure 4, was then compared with cut scores derived from the Kilgus et al. (2013) investigation. This comparison permitted an evaluation of the extent to which the four-profile model conformed to the hypothesized risk-based model for SABRS use (see Figure 1). In accordance with hypotheses, the average student was (a) not at-risk on both SB (>12) and AB (>11) within Profile 1, (b) at-risk on SB (≤12) but not at-risk on AB (>11) within Profile 2, (c) not at-risk on SB (>12) but are at-risk on AB (≤11) within Profile 3, and at-risk on both SB (≤12) and AB (≤11) within Profile 4.

Discussion

In accordance with recommendations from Kane (2013), previous SABRS-related research has informed the development of an interpretation/use argument, defining distinct models for interpretation and use. A model for interpretation specified that each SABRS item is related to one of two narrow factors: Social Behavior (SB) or Academic Behavior (AB). The model further specified that each item is related to a broad factor indicative of General Behavior (GB). Two models for use were proposed. The first corresponded to a two-profile model, which indicated that the SABRS could be used to differentiate between those who were (a) at-risk on GB or (b) not at-risk on GB. The second

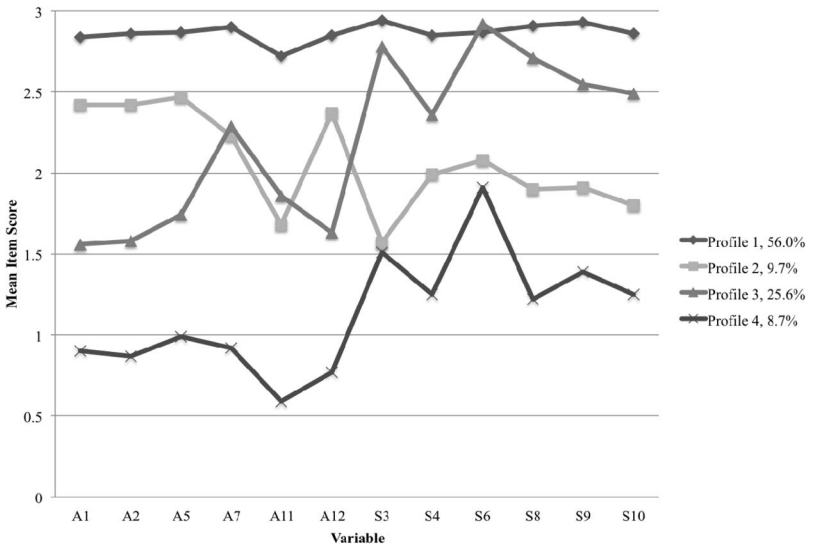

Figure 3. Four latent profiles corresponding to the Social and Academic Behavior Risk Screener (SABRS).

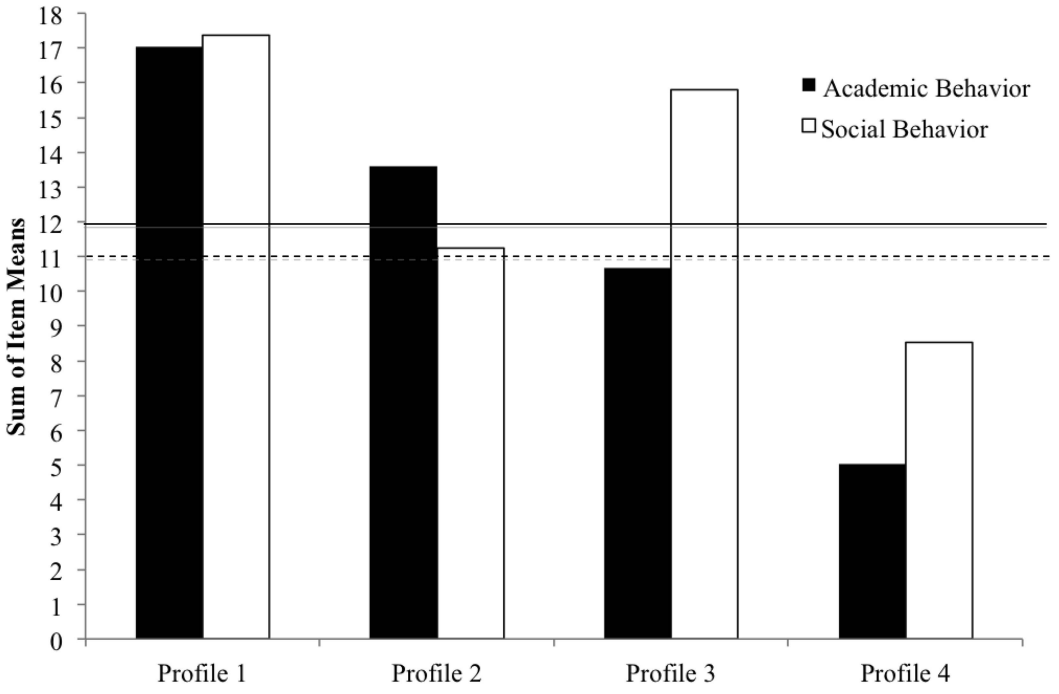

Figure 4. Sum of item means within each Social and Academic Behavior Risk Screener (SABRS) factor. Horizontal lines correspond to cut scores for the Social Behavior (≤12; solid line) and Academic Behavior (≤11; dashed line) factors.

model corresponded to a four-profile model, specifying that the SABRS could differentiate between those who were (a) not at-risk on either SB or AB, (b) not at-risk on SB but at-risk on AB, (c) at-risk on SB but not at-risk on AB, and (d) at-risk on both SB and AB. The purpose of the current study was to examine the tenability of these proposed models for interpretation and use.

Model for SABRS Interpretation

Factor and reliability analyses supported the proposed model for interpretation. Specifically, CFAs indicated that a bifactor structure yielded adequate fit that was superior to more parsimonious but conceptually consistent models. As previously noted, the bifactor model specified that all SABRS items were related to both a broad GB factor, as well as narrow SB and AB factors. More specifically, the model specified that the GB factor accounted for covariance among all items, and that the SB and AB factors accounted for residual covariance within item clusters (after controlling for GB). The bifactor structure is in accordance within the conceptual models that originally informed SABRS construction (Kilgus et al., 2013). Walker et al. (1992) proposed that a series of adaptive and maladaptive behaviors influence a student's propensity for social-behavioral competence. Kilgus and colleagues' (2013) review of the model indicated that although all behaviors are assumed related to this common outcome, one could argue that each falls into one of two categories corresponding to "Social Behavior" or "Academic Behavior." In other words, though behaviors are primarily grouped into a single broader category indicative of social-behavioral competence, items might also be secondarily grouped into narrower categories representative of behavioral response classes. This structure is supported within the current investigation, indicating that although all items might be interpreted relative to a single broad factor, they might also be interpreted relative to narrow factors that account for variance over

and above that which was explained by the general factor.

The model for interpretation was further supported by reliability findings. Results were indicative of the internal consistency of the SB, AB, and GB scales, with all alpha coefficients falling in the acceptable range (Cortina, 1993). This finding aligns with previous research, which also demonstrated the high internal consistency of all scales (Kilgus et al., 2013). Existing evidence suggests that despite a relatively limited number of items, the large degree of interitem correlation within each cluster supports the reliability of each scale and its interpretation in accordance with the proposed model. Such interpretation was further supported by interrater reliability findings. Although results seemingly indicated a degree of unreliability across raters, as ICCs, Pearson's *r* coefficients, and κ values did not exceed the "fair" range, a review of the literature indicated the current interrater reliability estimates were similar to if not slightly greater than the reliabilities associated with alternative universal screeners. Lane, Kalberg, Parks, and Carter (2008) reported a mean interrater reliability Pearson's *r* of .39 for the SRSS, with values ranging between .19 and .50. King et al. (2012) reported similar findings for the BESS (Mean $r = .30$, Range = .11–.39), as did Stone, Otten, Engels, Vermulst, and Janssens (2001) for the *Strengths and Difficulties Questionnaire* (Mean $r = .36$, Range = .26–.47).

One could argue that such performance is to be expected of behavioral screeners. A fundamental assumption of applied behavior analysis pertains to the deterministic relationship between the environment and human behavior. That is, student behavior is expected to differ across settings in accordance with environmental variation. Therefore, we should anticipate raters would offer differing perspectives of student behavior if the settings within which they have interacted with students vary. Within the current investigation, differences were noted across raters in terms of the settings (e.g., structured vs. unstructured), instructional formats (e.g., lecture vs. lab), and time of day (e.g., morning vs. afternoon) in which teacher–student interactions took place. A finding of only moderate or fair agreement across raters should therefore be expected and considered encouraging. If findings had reflected greater agreement, one might unfortunately conclude that the SABRS was insensitive to variation in behavior. That being said, it is difficult to determine the extent to which reliability estimates reflect true variation in behavior across settings. Future research should therefore employ alternative gold standard screeners to permit a comparison of interrater reliability across multiple measures.

Implications of Bifactor Structure

As noted above, reliability and CFA findings support consideration of a bifactor structure, allowing for interpretation of SABRS items relative to both a broad factor and one of two orthogonal narrow factors (Reise, 2012). Interpretation of scores on the broad factor would likely be useful for schools interested in examining a student's overall risk for behavioral difficulty (DiStefano et al., 2013). Consideration of scores on the narrow factors would be relevant for schools interested in gaining a more detailed understanding of student risk, with the intent to determine which interventions might be appropriate for each student given the nature of his or her risk (see the 'Model for SABRS Use' section below for additional information regarding such decisions).

The fit of the bifactor structure suggests it would be appropriate to calculate and interpret latent trait scores as summary estimates within each of these factors (DeMars, 2013). Yet, difficulties are noted with the applied interpretation and use of bifactor model latent trait scores (DiStefano et al., 2013). First, many researchers and practitioners may lack basic knowledge required to appropriately interpret bifactor trait scores. That is, many may not know to interpret the narrow trait scores as estimates of residual item variance unexplained by the general factor (Reise, 2012). Second, calculation of such trait scores is likely to be difficult within applied settings, as it would require educators to employ sophisticated and potentially expensive scoring software. As a potential result of these concerns (among others), authors of alternative behavior screeners characterized by a bifactor structure continue to recommend applied interpretation and use of their screener's in lieu of latent trait scores. For example, although findings have supported the bifactor structure of the BESS Teacher Rating Scale for Preschoolers, use remains founded

upon norm-referenced *t* scores (DiStefano et al., 2013). It is likely such score calculation and interpretation would remain appropriate for the SABRS. This conclusion is supported by findings from Kilgus et al. (2013), which were indicative of the psychometric defensibility of summed scale scores.

Model for SABRS Use

Beyond the model for interpretation, the current findings also supported hypotheses regarding the proposed model for use. LPA results indicated that the four-profile model yielded superior fit relative to alternative two- and three-profile models. Furthermore, examination of mean SB and AB scale scores relative to SABRS cut scores identified via previous research (Kilgus et al., 2013) indicated that one could define each profile in terms of its students' risk for either social or academic behavior problems. Specifically, (a) Profile 1 corresponded to the absence of risk on both SB and AB, (b) Profile 2 to risk on SB but not on AB, (c) Profile 3 to risk on AB but not on SB, and (d) Profile 4 to risk on both SB and AB. In other words, the current findings indicated that not only did the hypothesized four-profile model fit best, but that the profile organization also conformed to the a priori–specified at-risk/not at-risk structure that is at the foundation of the SABRS model for use. Results therefore represent additional support for the applied use of previously identified cut scores in determining which students are at-risk for behavioral difficulty and therefore in need of intervention.

LPA findings support a more complex and nuanced use of SABRS data, wherein students can be identified as at-risk for either social behavior problems or academic behavior problems. This finding indicates that the SABRS may be used to fulfill the basic purpose of universal screening, which is to determine whether a student is at-risk or not for behavioral problems. However, the current findings also indicate that SABRS functionality extends beyond this basic level, as it might also be used determine the nature of a student's risk. Information regarding the nature of behavioral risk might be useful in deriving initial recommendations regarding the type of intervention a student would require to be successful in the school setting. Students at-risk in either area might benefit from antecedent and consequence strategies, such as those packaged through Check In/Check Out, in supporting the display of learned but underperformed behaviors. In addition, information regarding the nature of each student's risk might provide information regarding which skills should be instructed improve behavioral functioning. Students at-risk for social behavior problems might require targeted direct instruction of social skills to remediate social skill acquisition deficits. Similarly, those at-risk for academic behavior problems might require targeted direct instruction of academic enablers to remediate academic enabler acquisition deficits. Finally, students at-risk in both areas would likely require some more intensive form of intervention, as specified through problem identification assessment indicative of behavioral topography and function. Though this assessment process would be time and resource intensive, recent research has supported the use of brief problem identification procedures for all students, even those exhibiting moderate risk and assigned to receive Tier 2 intervention (Reinke, Stormont, Clare, Latimore, & Herman, 2013). Overall, although the collection of additional data would be necessary to corroborate each of the aforementioned decisions, such as part of a multiple gating procedure, information gained via the SABRS might permit expedition of assessment and intervention processes.

A notable LPA result pertains to the percentage of students within each profile. As mentioned above, cross-referencing of previously specified SABRS cut scores and current LPA findings indicates that within the current sample, 56% of students fell in Profile 1 (i.e., not at-risk on either SB or AB), 36% in Profiles 2 and 3 (i.e., at-risk on either SB or AB, but not both), and 9% in Profile 4 (at-risk on both SB and AB). With approximately 45% of students falling in profiles characterized by risk within one or both domains, one might conclude that these findings represent a departure from widely recognized population-based estimates of behavioral risk (i.e., 20%; Schanding & Nowell, 2013), and thus potentially call into question SABRS performance within this study. However, we caution the reader against deriving such an interpretation. LPA represented a model-based approach to evaluating whether the SABRS evidenced unique profiles of student functioning. That is, findings were indicative of the SABRS' ability to differentiate between at-risk and not at-risk students across multiple narrow factors. Though LPA indicates the percentage of student within each profile, these findings are

not to be interpreted as estimates of the true prevalence of behavioral risk within the current student sample. Given LPA support for the use of SABRS to differentiate risk across multiple factors, a second step involves the specification of cut scores to determine the percent of students falling below each cut score. A review of the data indicated that in contrast to LPA-based percentages, only 15% and 24% of students fell below cut scores for SB and AB, respectively. Such estimates are considered to be in accordance with recent population-based estimates, suggesting the concordance between the SABRS and alternative screening approaches (e.g., Behavioral and Emotional Screening System; Schanding & Nowell, 2013).

Limitations

Multiple limitations to the current findings should be noted. First, the generalizability of the results is somewhat limited, as teachers and students were sampled from a single Midwestern high school. Furthermore, both teachers and students were predominantly White, indicating a large degree of homogeneity. As such, it is unlikely that the current findings generalize to the broader United States population, which is characterized by a much larger degree of ethnic diversity. Future investigations should look to employ larger and more diverse samples across multiple schools, districts, and geographic areas. Diverse samples would further allow for the evaluation of SABRS bias and fairness, or the extent to which SABRS items function similarly across various ethnic groups. Collection of such evidence is considered highly necessary in justifying applied use of the SABRS within universal screening. Second, the manner in which teachers were selected to rate each student was nonsystematic and voluntary in nature. Future research should look to employ a more rigorous approach, such as random selection of at least three teacher raters for each student. Third, there is an absence of information regarding the integrity of assessment procedures. We are therefore unable to determine the extent to which teachers completed the SABRS in accordance with researcher recommendations. We are also unable to corroborate expectations regarding the amount of time required to complete the SABRS for each student (i.e., 2–3 minutes). Future investigations should look to collect information regarding assessment integrity, as well as data regarding the efficiency of SABRS procedures, as each speaks to the ultimate usability of the measure in universal screening.

Future Directions for Research

The previously described interpretation/use argument defines a roadmap for future SABRS-related research (Kane, 2001, 2013), specifying what evidence is necessary to justify applied use of the SABRS for universal screening purposes. Research regarding the model for interpretation should proceed along multiple lines. First, investigators should look to corroborate the appropriateness of the bifactor model demonstrated herein. Although an initial EFA was conducted at the elementary level, the current high school–based study is the first to conduct a CFA of the proposed model for interpretation. Additional research should therefore consider the bifactor model via CFA at the elementary and middle school levels. Second, researchers should conduct additional examinations of internal consistency and interrater reliability, while also considering additional reliability types (e.g., test-retest). Related to this point, there is a need to consider the degree of consistency in SABRS data within a school year, because results have the potential to inform recommendations regarding the number of necessary annual screening administrations.

Additional research regarding the SABRS model for use should examine the screener's ability to identify the presence of risk, while also differentiating between different types of risk. Researchers should further examine the defensibility of SABRS cut scores (Kilgus et al., 2013). This may be accomplished via further application of LPA, as well as receiver operating characteristic (ROC) curve analysis. Such research should examine the extent to which SABRS cut scores differentiate between students on various gold standard criterion measures and key student outcomes, including school dropout, mental health diagnosis, grade retention, and special education placement. Consideration should be given to the necessity of varying cut scores across and within varying school grades in the interest of maximizing correct decision making.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317–332. doi:10.1007/BF02294359

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246. doi:10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606. doi:10.1037/0033-2909.88.3.588

Beyers, W., & Seiffge-Krenke, I. (2007). Are friends and romantic partners the "best medicine?" How the quality of other close relations mediates the impact of changing family relationships on adjustment. *International Journal of Behavioral Development, 31,* 559–568. doi:10.1177/0165025407080583

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290. doi:10.1037/1040-3590.6.4.284

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

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16–29. doi:10.1037/1082-989X.1.1.16

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13,* 354–378. doi:10.1080/15305058.2013.799067

DiStefano, C., Greer, F. W., & Kamphaus, R. W. (2013). Multifactor modeling of emotional and behavioral risk of preschool-age children. *Psychological Assessment, 25,* 467–476. doi:10.1037/a0031393

Drummond, T. (1994). *The student risk screening scale (SRSS)*. Grants Pass, OR: Josephine County Mental Health Program.

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

Forstmeier, S., & Maercker, A. (2007). Comparison of two diagnostic systems for complicated grief. *Journal of Affect Disorders, 99,* 203–211. doi:10.1016/j.jad.2006.09.013

Herman, K. C., Ostrander, R., Walkup, J. T., Silva, S. G., & March, J. S. (2007). Empirically derived subtypes of adolescent depression: Latent profile analysis of co-occurring symptoms in the treatment for adolescents with depression study (TADS). *Journal of Consulting and Clinical Psychology, 75,* 716–728. doi:10.1037/0022-006X.75.5.716

Hix-Small, H., Duncan, T. E., Duncan, S. C., & Okut, H. (2004). A multivariate associative finite growth mixture modeling approach examining adolescent alcohol and marijuana use. *Journal of Psychopathology and Behavioral Assessment, 26,* 255–270. doi:10.1023/B:JOBA.0000045341.56296.fa

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

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.

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.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2013). Validating the interpretations and uses of tests scores. *Journal of Educational Measurement, 50,* 1–73. doi:10.1111/jedm.12000

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. doi:10.1037/spq0000024

King, K., Reschly, A. L., & Appleton, J. J. (2012). An examination of the validity of the Behavioral and Emotional Screening System in a rural elementary school. *Journal of Psychoeducational Assessment, 30,* 527–538. doi:10.1177/0734282912440673

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

Lane, K. L., Kalberg, J. R., Parks, R. J., & Carter, E. W. (2008). Student Risk Screening Scale: Initial evidence for score reliability and validity at the high school level. *Journal of Emotional and Behavioral Disorders, 16,* 178–190. doi:10.1177/1063426608314218

Lane, K. L., 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. doi:10.1177/1063426608326203

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, NY: Guilford Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Levitt, J., Saka, N., Romanelli, L. H., & Hoagwood, K. (2007). Early identification of mental health

problems in schools: The status of instrumentation. *Journal of School Psychology, 45,* 163–191. doi:10.1016/j.jsp.2006.11.005

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99. doi:10.1037/1082-989X.4.1.84

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. doi:10.1037/0012-1649.41.5.733

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley. doi:10.1002/0471721182

Muthén, L. K., & Muthén, B. O. (1998–2013). *Mplus* (Version 7.1). Los Angeles, CA: Author.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

National Research Council and Institute of Medicine. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities* (M. E. O'Connell, T. Boat, & K. E. Warner, Eds.). Washington, DC: National Academies Press.

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

Nylund, K. L., Asparouhov, A., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14,* 535–569. doi:10.1080/10705510701575396

Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology, 32,* 8–47. doi:10.1016/j.cedpsych.2006.10.003

Reinke, W. M., Stormont, M., Clare, A., Latimore, T., & Herman, K. C. (2013). Differentiating tier 2 social behavioral interventions according to function of behavior. *Journal of Applied School Psychology, 29,* 148–166. doi:10.1080/15377903.2013.778771

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696. doi:10.1080/00273171.2012.715555

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley. doi:10.1002/9780470316696

Samuelsen, K., & Raczynski, K. (2013). Latent class/profile analysis. In Y. Petscher, C. Schatschneider, & D. L. Compton (Ed.), *Applied quantitative analysis in education and the social sciences* (pp. 304–328). New York, NY: Routledge.

Schanding, G. T., & Nowell, K. J. (2013). Universal screening for emotional and behavioral problems: Fitting a population-based model. *Journal of Applied School Psychology, 29,* 104–119. doi:10.1080/15377903.2013.751479

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

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

Steiger, J. H., & Lind, J. C. (1980, June). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Stone, L. L., Otten, R., Engels, R. C. M. E., Vermulst, A. A., & Janssens, J. M. A. M. (2001). Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: A review. *Clinical Child and Family Psychology Review, 13,* 254–274. doi:10.1007/s10567-010-0071-2

Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in a growth mixture models. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Greenwich, CT: Information Age.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10. doi:10.1007/BF02291170

von der Embse, N., Mata, A., Segool, N., & Scott, E. C. (2014). Latent profile analysis of test anxiety: A pilot study. *Journal of Psychoeducational Assessment, 32,* 165–172. doi:10.1177/0734282913504541

Walker, H. M., Irvin, L. K., Noell, J., & Singer, G. H. S. (1992). A construct score approach to the assessment of social competence: Rationale, technological considerations, and anticipated outcomes. *Behavior Modification, 16,* 448–474. doi:10.1177/01454455920164002

Received January 28, 2014
Revision received July 18, 2014
Accepted July 28, 2014 ■