11-6-2016

Latent Structure of Scores From the Emotional and Behavioral Screener

Matthew C. Lambert
University of Nebraska-Lincoln, mlambert2@unl.edu

Stacy-Ann A. January
University of South Carolina, sajanuary@sc.edu

Corey D Pierce
University of Northern Colorado, Corey.Pierce@unco.edu

Published in *Journal of Psychoeducational Assessment,* 12 pp. doi: 10.1177/0734282916676130

Published online November 6, 2016.

Latent Structure of Scores From the Emotional and Behavioral Screener

Matthew C. Lambert,[1] Stacy-Ann A. January,[2] and Corey D. Pierce[3]

1 University of Nebraska–Lincoln
2 University of South Carolina, Columbia
3 University of Northern Colorado, Greeley

Corresponding author — Matthew C. Lambert, Department of Special Education and Communication Disorders, University of Nebraska– Lincoln, 273 Barkley Memorial Center, Lincoln, NE 68583, USA. Email: mlambert2@unl.edu

Abstract

The Emotional and Behavioral Screener (EBS) is a recently developed teacher-reported brief screening instrument for identifying students who are at-risk of an emotional or behavioral disorder (EBD). Although prior research supports the technical adequacy of scores from the EBS, there is a gap in the literature regarding strong evidence of the factor structure underlying EBS scores. This study investigated the latent structure of scores from the EBS in a sample of 646 elementary students who were rated by their teachers in a 2-week screening period. Single-factor confirmatory factor analysis (CFA) and bifactor models were used to test the hypothesis that EBS scores are a measure of both overall emotional and behavioral risk and students' externalizing and internalizing behaviors. Results supported a bifactor structure, in that scores from the EBS can be considered to represent a general factor (i.e., risk of EBD) and two group factors (i.e., externalizing and internalizing domains). Findings have implications for interpreting scores when using the EBS as a universal screener for the risk of EBD.

Keywords: emotional and behavioral screener, factor structure, behavior rating scale, universal screening, prevention

Universal screening within prevention-oriented frameworks, such as Multi-Tiered Systems of Support (MTSS), is one way in which schools can promote the academic and behavioral well-being of all students (Dowdy, Ritchey, & Kamphaus, 2010; Sugai & Horner, 2009). Data from screenings can be used within a multiple gating approach in which at-risk students are followed up with additional assessment before receiving services targeted to reduce risk and/or the development of a disability (Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007). Currently, many more schools conduct universal screenings for academic difficulties than those that engage in screening in the area of mental health (Bruhn, Woods-Groves, & Huddle, 2014; Romer & McIntosh, 2005). Evidence that approximately 20% of children and adolescents have a mental health disorder that may result in impairment in school and other settings (Costello, Mustillo, Erkanli, Keeler, & Angold, 2003; Merikangas et al., 2010) suggests that students with emotional and behavioral difficulties are underidentified in schools. Furthermore, students who develop a diagnosable disorder often exhibit early indicators (Pihlakoski et al., 2006) that could be identified through systematic screening for emotional and behavioral risk.

Emotional and behavioral risk can be defined as the early symptoms of behaviors that make students more susceptible to later developing an emotional or behavioral disorder (EBD; Kamphaus, 2012). Compared with those without disabilities, students with EBDs have poor outcomes (Bradley, Doolittle, & Bartolotta, 2008). That is, students with EBD score lower on measures of academic achievement, are suspended at high rates, are often absent from school, and are more likely to drop out of high school (Wagner, 1995; Wagner, Kutash, Duchnowski, Epstein, & Sumi, 2005). Once they leave school, students with EBD have high rates of unemployment and involvement in the justice system, and are less likely to enroll in postsecondary education (Wagner et al., 2005). Given the negative outcomes of students with EBD, schools should conduct systematic screening for emotional and behavioral risk, as it is a more reliable method (as compared with teacher or self-referral) for identifying students who are at risk for mental health disorders (Eklund et al., 2009; Husky et al., 2011).

In addition to being technically adequate and easy to use (Glover & Albers, 2007), Emotional and Behavioral Screeners (EBSs) must sufficiently measure the construct of interest with appropriate theoretical and empirical support (Glover & Albers, 2007; Kane, 2013). Within the context of school-based prevention of EBD, screeners should measure early symptoms that are predictive of a range of negative mental health outcomes (Kamphaus, Reynolds, & Dever, 2014; National Research Council and Institute of Medicine Committee on the Prevention of Mental Disorders and Substance Abuse Among Children, Youth and Young Adults: Research Advances and Promising Interventions, 2009). This implies that screeners should capture indicators of externalizing behavior problems (e.g., aggression) and internalizing problems (e.g., anxiety), as these are constructs that are related to mental health. This is particularly salient within the school context, as reliance solely on teacher referral often results in a focus on externalizing problems, as students with internalizing problems are often underreferred (McIntosh, Ty, & Miller, 2014). Within the context of school-based prevention, it is relevant and important that screening measures include indicators of the characteristics of the federal definition of EBD.

EBS

One recently developed universal screening measure is the EBS (Cullinan & Epstein, 2013b). The EBS is a teacher report universal screener designed to assess the early indication of symptoms that would be predictive of EBD. The EBS was developed in response to the need for a brief, useful, and technically adequate universal behavioral screener to support schools' prevention-oriented systems (e.g., MTSS) that were adopted after substantial changes in federal legislation (i.e., 2004 reauthorization of the Individuals With Disabilities Education Improvement Act) that, among other things, called for the use of valid assessments for identifying students at risk of disabilities. The 10 items on the EBS were drawn from the Scales for Assessing Emotional Disturbance–Second Edition (SAED-2; Epstein & Cullinan, 2010), a nationally norm-referenced measure of the five characteristics of EBD. EBS items were drawn from the SAED-2 because it measures the five characteristics of emotional disturbance (ED) in the federal definition, is psychometrically sound, and has national norms (Epstein & Cullinan, 2010). During the development of the EBS, items were selected that represented each of the five characteristics of EBD and multiple trial versions with differing scale lengths were tested to examine how well the items discriminated between students with EBD and those without EBD in the SAED-2 normative sample. Given that each trial version performed similarly, the developers chose the version with the fewest (i.e., 10) items (Cullinan & Epstein, 2013b).

Initial studies examining the reliability and validity of scores from the EBS provide strong evidence of their psychometric properties. That is, EBS scores have strong internal consistency for students ages 5 to 11 years (α = .89) and 12 to 17 years (α = .90), and across race/ethnicity (r = .87.93) and gender (r = .89) groups (Cullinan & Epstein, 2013a). In addition, research supports the interrater reliability (r = .63), test–retest reliability (r = .90), social validity, and the convergent validity (r = .79, .87) of scores from the EBS (Nordness, Epstein, Cullinan, & Pierce, 2014; Pierce, Lambert, & Alamer, 2016). EBS scores also demonstrate diagnostic utility in accurately classifying students with risk of EBD across age (5-17; 86%-95% classification accuracy) and gender (87% classification accuracy for females; 86%-88% classification for males) groups (Lambert, Epstein, & Cullinan, 2014; Pierce et al., 2016).

Consistent with its intended use for screening within an MTSS framework, the EBS produces an overall score that is indicative of risk for EBD. Likewise, given the importance of adequate construct representation in a screener (Glover & Albers, 2007; Kane, 2013), the EBS was also designed to assess indicators of externalizing problems (e.g., aggression, impulsivity) and internalizing difficulties (e.g., anxiety, withdrawal), given the range of difficulties students might experience. That is, items on EBS were designed to reflect a hierarchical structure in which scores are indicators of risk for EBD while assessing indicators of both externalizing *and* internalizing problems, as these are the constructs underlying mental health (Kamphaus et al., 2014).

In an effort to support the construct validity of EBS scores, Lambert, Epstein, Ingram, Simpson, and Bernstein (2014) investigated the factor structure of scores from the EBS with a sample of first-grade students. They found some support for the internal structure of the EBS scores, but the confirmatory factor analysis (CFA) model demonstrated relatively poor fit to single-factor solution. Although the researchers suggested that the fit was adequate enough to assume a unidimensional structure based on the comparative fit index (CFI) and Tucker–Lewis index (TLI), the root mean square error of approximation (RMSEA) fit index was greater than .10, which is widely regarded as an indicator of poor model fit (Hu & Bentler, 1999). Therefore, evidence of the factor structure and reliability of the items comprising the EBS is unconvincing. Moreover, a focus on the underlying factor structure of a mental health screener is essential for establishing its appropriateness for the intended purpose (Kane, 2013) and can help inform research and practice. That is, without strong support for the interpretations of the EBS (i.e., that it is both an indicator of overall emotional and behavioral risk and externalizing and internalizing difficulties) practitioners cannot confidently use it to help make data-based decisions. This is a necessary prerequisite for understanding how well the EBS classifies who is at risk and who is not at risk.

Current Study

Although one previous study reported evidence of the factor structure and reliability of the EBS scores (e.g., Lambert, Epstein, Ingram, et al., 2014), the factor model goodness-of-fit was poor and the study only included students in first grade. The importance of establishing a tenable measurement model for the EBS scores cannot be overstated because the measurement model serves as the “basis and rationale for arriving at the composite [scores]” (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1999, p. 20), and is a prerequisite for assessing score reliability (Slaney & Maraun, 2008). Therefore, the purpose of this study was to examine the latent structure of EBS ratings using factor analysis approaches as well as to estimate the reliability of the composite scores for a sample of elementary school students.

Method

Participants

Participants included 626 elementary students in Grades 1 to 6 rated by 37 teachers from three schools within a large school district in the Rocky Mountain region. Students ranged in age from 5 to 12 years with an average age of 8.62 years (*SD* = 1.74 years). The gender composition of the sample was similar to the national population of school-age children with slightly more females (51%) than males (49%). Most students were White (57%) or Hispanic/Latino (39%) and a majority (63%) received free or reduced-price meals. Approximately 12% of students were English learners. Each of the 37 teachers was female and teaching in an elementary (K-5) classroom at the time of the study. The teachers had a wide range of teaching experience ranging from 2 to 29 years with an average of 11.97 years teaching experience. Of the 34 teachers who reported licensure information, 32 held an elementary teacher license and two held early childhood (birth to third grade) licensure in the state in which the study was conducted.

Measure

The EBS (Cullinan & Epstein, 2013b) is a 10-item teacher-rated scale used for screening students to identify which students may be at risk of EBD. A teacher or other adult familiar with the student completes the EBS form by rating each item on a 4-point Likert-type scale (0 = *not a problem*, 1 = *mild problem*, 2 = *considerable problem*, 3 = *severe problem*). The 10-item ratings are summed to form a Total EBS Score (score range = 0-30), which is compared with a cutoff score listed on the EBS form. Consistent with evidence that approximately 20% of students have a significant mental health problem (U.S. Department of Health and Human Services, 1999), the developers chose the 80th percentile as the cutoff point to indicate risk of EBD (Cullinan & Epstein, 2013a). A student who's *Total EBS Score* exceeds the cutoff score for their age and gender is identified as "at risk." The cutoff score for younger male students (5-11) is 5, younger female students (5-11) is 3, older male students (12-17) is 7, and older female students (12-17) is 4. The psychometric properties and diagnostic utility of the EBS scores are well established (Cullinan & Epstein, 2013a, 2013b; Lambert, Epstein, & Cullinan, 2014; Lambert, Epstein, Ingram, et al., 2014; Nordness et al., 2014; Pierce et al., 2016).

Procedures

The 37 teachers who participated in this study were asked to complete an EBS on each student in their homeroom classrooms. They were asked to complete the forms only on students who were not currently receiving special educational services for EBDs. A common identification number was used to link the two sets of scores. Each teacher was given a packet that contained enough EBS forms to rate each child in their classroom. Instructions for completing the forms were included in the packet. The teachers also completed a brief survey that asked them to provide demographic information on themselves (e.g., gender, teaching experience, licensure). Upon completing the ratings, teachers submitted their completed materials to a designated research liaison at each school. The liaison then notified the researchers when all materials were completed. Each participating teacher received remuneration of US$100.

Analysis Plan

Mplus v7.11 (Muthen & Muthen, 1998-2014) was used to fit three types of factor analysis models: (a) a single-factor CFA model, (b) an exploratory bifactor model, and (c) a confirmatory bifactor model (i.e., nested-factor model; general-specific model). As items were measured on a 4-point rating scale, weighted least square with mean and variance adjustments (WLSMV; also called robust WLS) was used to fit the CFA models. Missing data were negligible (<1%) and excluded from the analysis on a pairwise-present basis as is default in Mplus when using the WLSMV estimator (Muthen & Muthen, 1998-2014).

Chi-square (χ^2), the CFI (Bentler, 1990), the TLI (Tucker & Lewis, 1973), and the RMSEA (Steiger & Lind, 1980) were used to assess model fit. Standard cutoffs for acceptable CFI, TLI, and RMSEA are values greater than or equal to .90 for CFI and TLI (Browne & Cudeck, 1992; Hu & Bentler, 1999) and values less than .08 for RMSEA (Hu & Bentler, 1999). The DIFFTEST chi-square difference test ($\Delta\chi^2$) was used to evaluate the improvement in goodness-of-fit for the bifactor model over the single-factor model.

Bifactor models. The bifactor model assumes that each item response has two sources of systematic variance: (a) the general factor and (b) group factors (i.e., more narrow, specific constructs). That is, the rating for each EBS item is a function of the student's risk of EBD and the student's specific level of aggression, social isolation, impulsiveness, and so on. The bifactor model lends itself well to behavioral assessments because common "best" practice of assessment development is to include content heterogeneity when measuring behavioral constructs (Reise, 2012). This helps establish better content and construct validity, but can cause measurement bias when scores are used as if they are unidimensional. Assessments with any degree of content heterogeneity are by definition multidimensional (Harrison, 1986); however, the pertinent question is, are the scores "unidimensional enough" to be treated as such?

The findings from a bifactor model can be used to address the question of unidimensionality by determining the degree to which raw composite scores reflect a single common source of variability. As the variances of item responses are partitioned into common sources, the proportion of common variance attributable to the general factor can be expressed as the explained common variance (ECV) which indicates the relative "unidimensionality" of the explained variance. ECV values closer to one indicate a greater degree of unidimensionality. The degree to which assessment composite scores represent a single common source of variance is expressed as the omega hierarchical reliability coefficient (i.e., precision of the raw scores).

Omega hierarchical (ω_h) reliability (McDonald, 1999; Zinbarg, Yovel, Revelle, & McDonald, 2006) was estimated based on the confirmatory bifactor model and calculated using Equation 1, where λ_{iGEN} is the loading for each item on the general factor, λ_{iGRP} is the loading for each item on its group factor, and θ_i^2 is the error variance for each item. As the equation indicates, the ω_h coefficient includes a single variance component related to the general factor in the numerator and the total variance of scores in the denominator. Like coefficient alpha, omega hierarchical is a proportion and ranges from 0 (completely unreliable) to 1 (perfectly reliable). See McDonald (1999); Brunner, Nagy, and Wilhelm (2012); or Reise (2012) for a more through explanation of calculating and interpreting omega reliability estimates using parameters from CFA models.

$$\omega_h = \frac{\left(\sum \lambda_{iGEN}\right)^2}{\left(\sum\lambda_{iGEN}\right)^2 + \left(\sum\lambda_{iGRP_1}\right)^2 + \left(\sum\lambda_{iGRP_p}\right)^2 + \sum\theta_i^2} \qquad (1)$$

Results

Goodness-of-fit indicators (i.e., chi-square, CFI, TLI, RMSEA) for the two CFA models are presented in Table 1. The chi-square difference test between the single-factor and bifactor models is also reported

Table 1. Confirmatory Factor Analysis Model Fit Indicators.

	$\chi^2_{(df)}$	$\Delta\chi^2$	CFI	TLI	RMSEA [90% CI]
Confirmatory single factor	$345.65_{(35)}$	—	0.958	0.946	0.119 [.108, .131]
Exploratory bifactor					
Two group factors	$50.63_{(18)}$	—	0.996	0.989	0.054 [.037, .072]
Three group factors	$12.59_{(11)}$	—	1.000	0.999	0.015 [.000, .046]
Confirmatory bifactor	$98.27_{(26)}$	192.31*	0.990	0.983	0.067 [.053, .081]

$\Delta\chi^2$ for the two confirmatory models was calculated using the DIFFTEST feature in Mplus (Muthen & Muthen, 1998-2014). The degrees of freedom for the difference tests were calculated as the difference in the number of degrees of freedom between the two models being compared. CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval.

*$p < .0001$.

in Table 1. Factor loadings for the two models are presented in Table 2. Note that there are two sets of loadings for the bifactor model; one set for the general factor and one set for the group factors.

Single-Factor Model

The unidimensional model did not fit the EBS data acceptably as indicated by the RMSEA index (.119) which was considerably larger than the general cutoff of .08. Following conventional guidelines, the fit of the single-factor model is considered poor—the EBS scores are *not* strictly unidimensional. However, all of the factor loadings were large (>.50; see Table 2) which seems to indicate a strong single factor, yet the fit of the model was poor. This was because the source of misfit in the single-factor model was the intercorrelated residual item variances that arise due to multidimensionality (i.e., content heterogeneity of the EBS items).

In an attempt to account for the multidimensionality of the EBS ratings, we made simple modifications to the single-factor model guided by modification indices (MODINDICES). We correlated the residual variances of Items 1 and 7 because this was the largest modification index (MI = 121.74) and these two items seem to represent internalizing problems (*anxious* and *lacks self-confidence*). When allowing the residual variances of Items 1 and 7 to correlate, model fit improved ($\chi^2_{(34)}$ = 233.44, CFI = 0.973, TLI = 0.964, RMSEA = 0.097 [0.085, 0.109]), but the fit is still considered poor, as indicated by the large RMSEA value.

Bifactor Models

Because the EBS does not purport to consist of subscales, there was no immediate information on the bifactor structure of the EBS scores (i.e., which items should load onto which *group* factors). Therefore, prior to fitting a confirmatory bifactor model, exploratory bifactor models were estimated using the exploratory factor analysis (EFA) functionality of Mplus with a bi-geomin rotation (Muthen & Muthen, 1998-2014) following recommendations by Reise, Moore, and Haviland (2010). A model with *two* group factors and a model with *three* group factors were extracted from the assessment data.

Exploratory bifactor model. The exploratory bifactor analyses indicated that both the two and three group factor models fit the data well (see Table 1). The loadings for the two group factor model indicated that items such as *anxious* (λ = 0.44), *lacks self-confidence* (λ = 0.70), and *lacks social skills* (λ = 0.20) factor together (this factor might represent internalizing or social problems) while most of the remaining items factor together (this factor might represent externalizing problems) with

Table 2. Factor Loadings From Confirmatory Factor Analyses.

		Confirmatory bifactor	
	Single factor	General	Group
Group factor—internalizing			
Anxious	.59	.57	.49
Lacks self-confidence	.55	.51	.66
Lacks social skills	.86	.90	.13
Group factor—externalizing			
Destroys things	.84	.73	.44
Disrespectful	.84	.65	.59
Does not work well in groups	.91	.90	.24
Fails to consider consequences	.90	.76	.56
Gets distracted	.74	.65	.38
Makes threats	.87	.74	.49
No group factor			
Rejected by peers	.90	.96	—

the exception of the item *rejected by peers* which only significantly loaded on the general factor. Although the three group factor solution represented the data very well, interpreting the meaning of the group factors was difficult. In addition, a CFA model based on the exploratory three group factor solution is not identified (i.e., there are more unknown model parameters than "pieces" of known information. Given the limitations of the three group factor solution, the two group factor model was used to develop the subsequent CFA model. It should be noted that group factors in the exploratory models are oblique (i.e., correlated) while group factors in the confirmatory model are orthogonal (i.e., uncorrelated). This difference in model specification can lead to substantial distortions in factor loadings and model fit between exploratory and confirmatory bifactor models (Finch, 2011).

Confirmatory bifactor model. The confirmatory bifactor model was fit with two group factors, representing internalizing and externalizing domains, and one general factor representing risk of EBD. Note that this model is a so-called "incomplete" bifactor model because not all items (i.e., *rejected by peers*) loaded onto a group factor (see Figure 1). The model fit the data acceptably (CFI = 0.990, TLI = 0.983, RMSEA = 0.067) and represented a statistically significant improvement in fit over the single-factor model as indicated by the chi-square difference test ($\Delta\chi^2_{(9)} = 192.31$, $p < .0001$). The ECV for the general factor was 0.74, indicating that 74% of the common variance was attributable to the general factor. This might suggest that the EBS scores might be "unidimensional enough" to treat as such when creating composite scores or when using EBS data in structural equation modeling or item response theory modeling applications (Reise et al., 2010).

Model-based reliability. Omega hierarchical was used to evaluate the precision in which the EBS scores represent the underlying risk of EBD—the degree to which the scale scores can be interpreted as a measure of a single factor. The omega hierarchical estimate indicates the proportion of scale score variance that is attributed to the general factor (i.e., true score) when accounting for the group factor variances. Omega hierarchical for the general factor was .83 suggesting that the EBS scores primarily reflect the single source of common variance and are reliable measures of the risk of emotional disturbance.

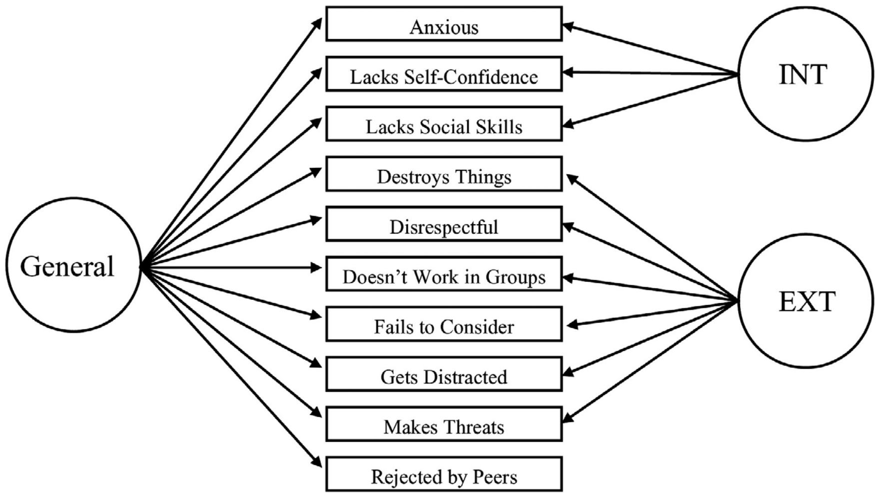

Figure 1. Bifactor model.

Discussion

The EBS is a recently developed brief screening instrument for identifying students who are at risk of EBD within MTSS frameworks. Although research supports the reliability and validity of EBS scores, strong evidence of the hierarchical latent structure of EBS scores has not yet been established. The purpose of this study was to (a) investigate the latent structure of scores from the EBS and (b) estimate the reliability of the composite scores with elementary students in Grades 1 through 6. Results of the factor analysis models supported a bifactor structure, indicating that scores from the EBS can be considered to represent a general factor and two group factors.

Findings from this study suggest that EBS scores can be considered univocal, even though a bifactor model fit statistically better than the single-factor model. The item-factor loadings for the single-factor model did not differ substantially from the loadings for the general factor, thus indicating that the same underlying factor is being measured by both models. In addition, the ECV estimate from the bifactor model suggested that the majority of systematic variance (74%) is attributable to the general factor. This finding is supported further by the high omega hierarchical reliability estimate (0.83) which indicated that the composite raw score primarily reflects the general factor variance. Therefore, results indicate that the overall score from the EBS reflects students' risk of EBD. Furthermore, by using EBS scores for decision making within MTSS frameworks, practitioners may help to reduce underidentification of elementary school students at risk of EBD.

Results from this study generally support the notion that externalizing and internalizing factors underlie scores from the EBS, suggesting that this 10-item measure captures the broad range of difficulties that are indicative of emotional and behavioral problems in school-age youth. That is, when compared with the single-factor model, loadings were almost always smaller for the bifactor model, providing some evidence of bias/distortion when creating composite scores based on a single-factor model. This slight distortion in scores is caused by content heterogeneity of the assessment items. This finding is encouraging, given that the EBS was designed to be a measure of overall risk as well as externalizing and internalizing symptomology. Notably, findings supported an "incomplete" bifactor model, as there was one item (i.e., rejected by peers) that did not load onto either of the two group factors of the confirmatory bifactor model. There are two potential explanations for this finding. First, empirical evidence and theory suggests that peer rejection is associated with both internalizing and externalizing behavior problems (Coie, Terry, Lenox, Lochman, & Hyman, 1995; Ladd, 2006; Schwartz, McFadyen-Ketchum, Dodge, Pettit, & Bates, 1998). Second, the rejected by peers item demonstrated a very high loading (λ = 0.96) onto the general factor. Thus, it could be that there was not enough residual variance left to be explained by one of the group factors. Nonetheless, findings from this study, combined with existing research (e.g., Cullinan & Epstein, 2013a; Lambert, Epstein, & Cullinan, 2014; Pierce et al., 2016) support the use scores from the EBS to identify students who are at risk of EBD.

Limitations

There are several potential limitations to this study that should be acknowledged. First, characteristics of this sample may limit the external validity (i.e., generalizability) of the findings of this study. For example, these data were collected from a single school district, and although the racial and ethnic diversity of the sample is reflective of region of the United States from which participants were drawn, there are differences between the demographics of the current sample and the national student population. Moreover, these data were all collected during one universal screening window. Thus, it is possible that the factor structure may differ for data collected during other time periods of the academic year. In addition, although the two factor bifactor model had the best fit to the data, that the *rejected by peers* item did not load onto one of the group factors could be considered a limitation because it means that the data do not strictly conform to the bifactor structure (i.e., every item has two sources of systematic variance). The item did not load unto either group factor because the variance of the item was nearly entirely explained by the general factor (R^2 = 92.16%) and the limited residual variance was random. Furthermore, the assessment data had an underlying nested structure (i.e., students nested within teachers); however, the CFA models did not account for the nesting due to the anonymity of data collection procedures (i.e., we did not record which students were nested within which teachers). By not accounting for the nested structure, there is potential for bias in CFA model parameters (e.g., item thresholds and factor loadings) and model fit due to a violation of conditional independence (Pornprasertmanit, Lee, & Preacher, 2014).

Future Research

Findings from the current study suggest several potential avenues for future research. For instance, it is possible that the factor structure of EBS scores could vary across age groups. Thus, future research should investigate the latent structure of scores from the EBS in older students (e.g., Grades 4–6) as compared with younger students (e.g., Grades K–3). In addition, although results of this study support the bifactor structure of EBS scores, research is needed to test the measurement invariance of the bifactor model parameters across a number of groups (e.g., gender, race/ethnicity). Finally, researchers might also investigate EBS ratings using probabilistic diagnostic models (e.g., latent class analysis) with a focus on comparing diagnostics model results with the current scoring method of the EBS.

Authors' Note — The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests — The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding — The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the Bresnahan-Halstead Center in the School of Special Education at the University of Northern Colorado. The development and preparation of this article was also supported by the Institute of Education Sciences (IES), U.S. Department of Education, through Grant R324B110001 to the University of Nebraska–Lincoln.

References

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

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

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

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230-258. doi:10.1177/0049124192021002005

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

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*, 796-846. doi:10.1111/j.1467-6494.2011.00749.x

Coie, J. D., Terry, R., Lenox, K., Lochman, J., & Hyman, C. (1995). Childhood peer rejection and aggression as predictors of stable patterns of adolescent disorder. *Development and Psychopathology*, *7*, 697- 713. doi:10.1017/S0954579400006799

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry*, *60*, 837-844. doi:10.1001/archpsyc.60.8.837

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

Cullinan, D., & Epstein, M. H. (2013b). *Emotional and Behavioral Screener*. Austin, TX: PRO-ED.

Dowdy, E., Ritchey, K., & Kamphaus, R. W. (2010). School-based screening: A population-based approach to inform and monitor children's mental health needs. *School Mental Health*, *2*, 166-176. doi:10.1007/s12310-010-9036-3

Eklund, K., Renshaw, T. L., Dowdy, E., Jimerson, S. R., Hart, S. R., Jones, C. N., & Earhart, J. (2009). Early identification of behavioral and emotional problems in youth: Universal screening versus teacher-referral identification. *The California School Psychologist*, *14*, 89-95. doi:10.1007/BF03340954

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

Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement*, *35*, 67-82. doi:10.1177/0146621610367787

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

Harrison, D. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, *11*, 91-115. doi:10.2307/1164972

Hu, L.-T., & 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

Husky, M. M., Kaplan, A., McGuire, L., Flynn, L., Chrostowski, C., & Olfson, M. (2011). Identifying adolescents at risk through voluntary school-based mental health screening. *Journal of Adolescence, 34*, 505-511. doi:10.1016/j.adolescence.2010.05.018

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., & Dever, B. V. (2014). Behavioral and mental health screening. In R. J.

Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools* (pp. 249-273). Washington, DC: American Psychological Association.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*, 448-457.

Ladd, G. W. (2006). Peer rejection, aggressive or withdrawn behavior, and psychological maladjustment from ages 5 to 12: An examination of four predictive models. *Child Development*, *77*, 822-846. doi:10.1111/j.1467-8624.2006.00905.x

Lambert, M. C., Epstein, M. H., & Cullinan, D. (2014). The diagnostic quality of the Emotional and Behavioral Screener. *Journal of Psychoeducational Assessment*, *32*, 51-61. doi:10.1177/0734282913485541

Lambert, M. C., Epstein, M. H., Ingram, S., Simpson, A., & Bernstein, S. (2014). Psychometrics and measurement invariance of the Emotional and Behavioral Screener. *Behavioral Disorders*, *39*, 89-101.

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

McIntosh, K., Ty, S. V., & Miller, L. D. (2014). Effects of school-wide positive behavioral interventions and supports on internalizing problems: Current evidence and future directions. *Journal of Positive Behavior Interventions*, *16*, 209-218. doi:10.1177/1098300713491980

Merikangas, K. R., He, J.-P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., ... Swendsen, J. (2010). Lifetime prevalence of mental disorders in U.S. adolescents: Results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 980-989. doi:10.1016/j.jaac.2010.05.017

Muthen, L. K., & Muthen, B. O. (1998-2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

National Research Council and Institute of Medicine Committee on the Prevention of Mental Disorders and Substance Abuse Among Children, Youth and Young Adults: Research Advances and Promising Interventions (with O'Connell, M. E., Boat, T., & Warner, K. E., Eds.). (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. Washington, DC: National Academies Press.

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

Pierce, C. D., Lambert, M. C., & Alamer, H. (2016). Convergent, criterion and social validity of the emotional and behavioral screener. *Journal of Child and Family Studies*, *25*, 77-85. doi:10.1007/s10826-015-0206-7

Pihlakoski, L., Sourander, A., Aromaa, M., Rautava, P., Helenius, H., & Sillanpaa, M. (2006). The continuity of psychopathology from early childhood to preadolescence: A prospective cohort study of 3-12-year-old children. *European Child & Adolescent Psychiatry*, *15*, 409-417. doi:10.1007/s00787-006-0548-1

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research*, *49*, 518-543. doi:10.1080/00273171.2014.933762

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

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544-559. doi:10.1080/00223891.2010.496477

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

Schwartz, D., McFadyen-Ketchum, S. A., Dodge, K. A., Pettit, G. S., & Bates, J. E. (1998). Peer group victimization as a predictor of children's behavior problems at home and in school. *Development and Psychopathology*, *10*, 87-99. doi:10.1017/S095457949800131X

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

Slaney, K. L., & Maraun, M. D. (2008). A proposed framework for conducting data-based test analysis. *Psychological Methods*, *13*, 376-390. doi:10.1037/a0014269

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.

Sugai, G., & Horner, R. H. (2009). Responsiveness-to-intervention and school-wide positive behavior supports: Integration of multi-tiered system approaches. *Exceptionality*, *17*, 223-237. doi:10.1080/09362830903235375

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

U.S. Department of Health and Human Services. (1999). *Mental health: A report of the Surgeon General*. Rockville, MD: Author.

Wagner, M., Kutash, K., Duchnowski, A. J., Epstein, M. H., & Sumi, W. C. (2005). The children and youth we serve: A national picture of the characteristics of students with Emotional Disturbances receiving special education. *Journal of Emotional and Behavioral Disorders*, *13*, 79-96. doi:10.1177/10634266 050130020201

Wagner, M. M. (1995). Outcomes for youths with serious emotional disturbance in secondary school and early adulthood. *Future of Children*, *5*, 90-112. doi:10.2307/1602359

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω_h. *Applied Psychological Measurement*, *30*, 121-144.