

Investigation of Factor Structure and Measurement Invariance by Gender for the Behavioral and Emotional Screening System Among High School Students

Jihye Kim
Kennesaw State University

Randy W. Kamphaus
University of Oregon

The Behavioral and Emotional Screening System (BESS; Kamphaus & Reynolds, 2007) is used by U.S. schools to screen students for behavioral and emotional risks. The BESS comprises a student form, a teacher form, and a parent form. We explored the factor structure of the BESS Student Form among high school students and evaluated measurement invariance by gender using multiple-group confirmatory factor analysis. The results suggested that the BESS Student Form has a 5-factor structure and holds to partial invariance by gender. The partial invariance model revealed 5 items that functioned differentially. These item-level findings suggest that female and male adolescents have different perspectives about their relationships with their parents, self-esteem, self-reliance, and atypicality. The implications of our study are twofold: The BESS Student Form might be used with female and male adolescents in its current form, but the analysis of the data suggested that possible changes be made in the item content during the next revision.

Public Significance Statement

This study demonstrates the difference of factor structures between high school students and elementary school students in the Behavioral and Emotional Screening System (BESS) Student Form and suggests that the BESS Student Form performs uniformly by gender.

Keywords: behavioral and emotional risk, measurement invariance, factor analysis, differential item functioning

Children with emotional and behavioral problems tend to show poor academic outcomes, including performance below grade in reading and mathematics (Bradley, Doolittle, & Bartolotta, 2008), higher rates of suspensions and expulsions (Wagner, Kutash, Duchnowski, & Epstein, 2005), and more frequent absenteeism (Lane, Carter, Pierson, & Glaeser, 2006). Considerable research has suggested that screening, along with early prevention and intervention strategies, can have a positive impact on children. For example, they are more likely to receive appropriate mental health or special education services, have better academic outcomes and pursue postsecondary education and are less likely to drop out of school (Atkins, Frazier, Adil,

& Talbott, 2003; Catalano, Haggerty, Oesterle, Fleming, & Hawkins, 2004; McIntosh, Flannery, Sugai, Braun, & Cochrane, 2008; Wagner et al., 2005). These positive impacts, however, are likely to happen only when coupled with the accurate identification of children with behavioral and emotional risks using well-validated instruments (Goe, Bell, & Little, 2008; Plecki, Elfers, & Nakamura, 2012).

Validity and reliability are the most fundamental elements of instrument validation. In addition, it is important to establish the degree to which the instrument performs uniformly across specific groups (e.g., female and male), also termed *measurement invariance* (MI). Differences in instrument performance across groups can lead to errors in screening individual participants, inaccurate estimates in research, and a lack of fairness and equity between groups (Millsap & Kwok, 2004).

The Behavioral and Emotional Screening System (BESS; Kamphaus & Reynolds, 2007) is one of the behavioral and emotional risk screening instruments widely used in school systems with children from preschool to high school. It consists of a student form, a teacher form, and a parent form. The BESS items were drawn from the Behavioral Assessment System for Children—Second Edition (BASC-2; Reynolds & Kamphaus, 2004), which assesses children's behavioral and emotional problems for an array of constructs, including internalizing and externalizing problems. These long-standing constructs have been documented by decades

This article was published Online First March 16, 2017.

Jihye Kim, Department of Secondary and Middle Grades Education, Kennesaw State University; Randy W. Kamphaus, Department of Special Education and Clinical Sciences, University of Oregon.

Randy W. Kamphaus wishes to disclose a potential conflict of interest in that he receives a portion of royalties on net sales of the Behavioral and Emotional Screening System forms.

Correspondence concerning this article should be addressed to Jihye Kim, Department of Secondary and Middle Grades Education, Kennesaw State University, 580 Parliament Garden Way NW, Mail Drop 0122, Kennesaw GA 30144. E-mail: jkim192@kennesaw.edu

of research (e.g., Edelbrock & Achenbach, 1980; Frick, Burns, & Kamphaus, 2009; D. R. Peterson, 1961; Quay, 1987).

Researchers have found that the BESS (Kamphaus & Reynolds, 2007) has yielded various types of reliability and validity (i.e., .92 internal consistency, .80–.91 test–retest stability, .71–.83 interrater reliability, .70–.80 sensitivity, and .87–.95 specificity) to detect children at behavioral and emotional risk (Dever, Mays, Kamphaus, & Dowdy, 2012). Researchers have focused mainly on identifying four latent factors (i.e., personal adjustment, inattention or hyperactivity, internalizing problems, and school problems) representing the behavioral and emotional problems of elementary students (Dowdy et al., 2011), and they have paid less attention to the MI of the BESS. In particular, researchers who have studied the BESS (e.g., Hyde, Mezulis, & Abramson, 2008) have found that girls who have mental health issues tend to be at more severe risk of having internalizing problems, whereas boys who have mental health issues tend to be more at risk of inattention or hyperactivity problems. To validate whether gender differences regarding behavioral and emotional problems is true or is the result of inaccurate detection, it is vital to determine whether the instrument performs consistently across groups or whether true gender differences exist.

In addition, the four latent factors of the BESS were established from data obtained from elementary students, with less attention being paid to high school students. Our hypothesis and theoretical background were that the latent factor structures might have been different because children and adolescents mature physically, cognitively, and socially. And different factor structures can distinguish the different traits of behavioral and emotional development between children and adolescents, regardless of participants' characteristics (e.g., age, race, sociogeographic factors). Consequently, we sought to determine whether the four latent factors also represented the behavioral and emotional problems of high school students. Then, we conducted MI testing of male and female adolescents to determine whether the factor structures or items on the BESS Student Form functioned fairly, regardless of gender group membership. Although prior BESS research has provided supportive evidence of validity and reliability, it was important to establish evidence that MI also provides a fair screening form, regardless of the participants' characteristics. Two research questions guided the study:

1. Are the factor structures of the BESS different between high school students and elementary school students?
2. Does the BESS Student Form yield MI across gender?

Method

Measure

The BESS Student Form (Kamphaus & Reynolds, 2007) comprises 30 items, each with a 4-point response scale ranging from 0 (*never*) to 3 (*almost always*) designed to detect children's risk levels for behavioral and emotional problems. The sum of the item raw scores is transformed to a total *T* score using a linear transformation, where *T* scores of 20–60 are classified as a normal level of risk, 61–70 as an elevated level of risk, and scores of 71 and higher as an extremely elevated level of risk. These classifi-

cation labels were determined by prepublication receiver operator characteristic curve analyses and in accordance with the distance of the scores from the norming sample mean (see Kamphaus & Reynolds, 2007, for a detailed description of these analyses and findings). The BESS yields only a single total score and score inference: Is behavioral and emotional risk present for this child or adolescent? These cutoff scores are intended to help practitioners to make decisions about which students might need additional assessment or services.

Data Collection

We used a nested data structure for students in Grades 9–12 in three high schools in the Los Angeles United School District (LAUSD) during the fall semester of their 2011–2012 academic year ($N = 4,017$). We used a passive parental consent protocol approved by the LAUSD to screen students in classroom groups in about half a school day at each high school. Researchers proctored the screening in each classroom using a set protocol that included obtaining student consent to participate in the screening. The students completed the BESS Student Form in a single class period. Only the English version was used in this study, even though we had offered the Spanish form to the students.

Of the 4,017 students, 45.5% ($n = 1,858$) were female. The grade-level sample sizes were 25.8% ($n = 1,035$) in Grade 9; 25.9% ($n = 1,042$) in Grade 10; 27.2% ($n = 1,094$) in Grade 11; and 21.1% ($n = 846$) in Grade 12. In total, about 88% ($n = 3,590$) obtained total *T* scores in the normal risk range, and 12% ($n = 484$) showed some level of risk (9% elevated and 3% extremely elevated; see Figure 1).

Analyses

Initial confirmatory factor analysis (CFA). Based on the four-factor structure of the student form established by Dowdy et al. (2011)—namely, personal adjustment, inattention or hyperactivity, internalizing problems, and school problems—we ran the initial CFA using high school students' BESS data to confirm whether the same factor structures for elementary school students existed among high school students (see Figure 2). Consistent with Dowdy et al., we replicated the study conditions of the initial CFA by using only 27 of the 30 items. The three items eliminated were Item 9 (being liked by others), Item 11 (difficulty sitting still), and Item 22 (feeling stupid). Of the 27 remaining items, two pairs of residuals were correlated: Item 29 (school comfort) with Item 12 (school interest) and Item 30 (others respect me) with Item 21 (others think I'm fun to be with). For evaluating the model fit, we considered the criteria of root-mean-square error of approximation (RMSEA) less than .08, comparative fit index (CFI) more than .90, Δ CFI .01 or less, and Tucker–Lewis index (TLI) more than .90 as acceptable fit indices (Cheung & Rensvold, 2002; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Garver & Mentzer, 1999).

Exploratory factor analysis (EFA). Because we found a poor model fit with the four-factor structure in the high school students' data from the initial CFA, we ran an EFA using all 30 items to find an optimal number of factors reflecting latent structures among high school students. We applied oblique promax rotation using Mplus (Version 7.2) [Computer software] and estimated EFA using the maximum likelihood estimation with robust

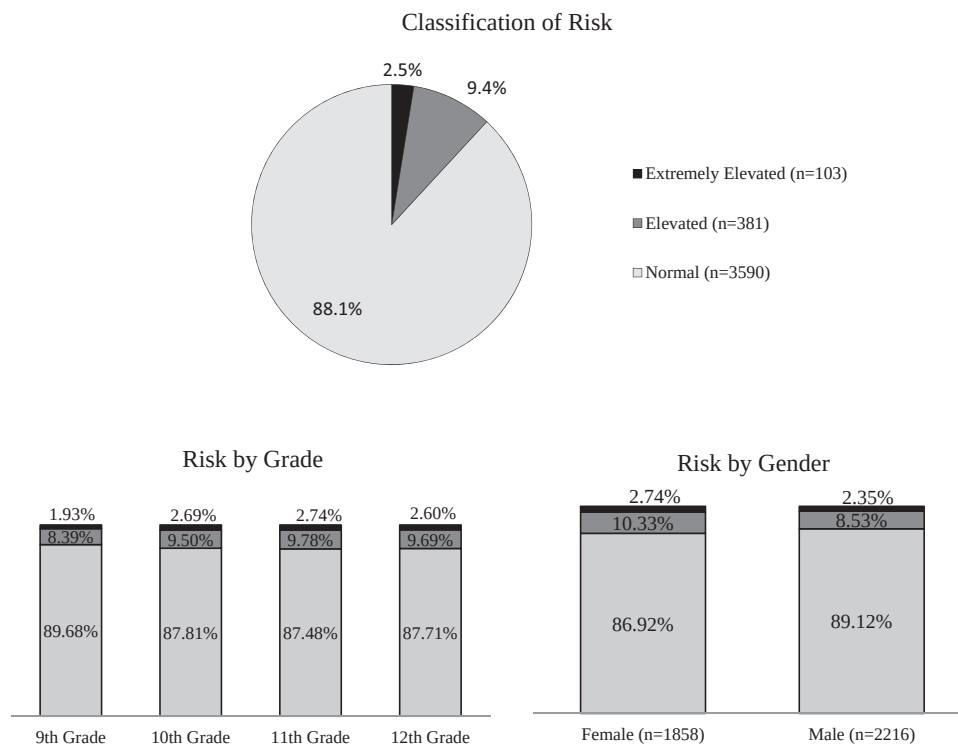


Figure 1. Classification of behavioral and emotional risk by grade and gender. Behavioral and emotional risk in BESS is categorized with three levels: normal, elevated, and extremely elevated.

standard errors (Costello & Osborne, 2005). We used scree plot, Horn's (1965) parallel analysis (PA), and several model fit indices (e.g., RMSEA, CFI, TLI) as factor retention methods. Horn's PA determines the number of factors to retain by generating random variables. The number of factors can be selected based on the percentile rank of expected eigenvalues from the simulation under the 95th percentile of the distribution (Cota, Longman, Holden, Fekken, & Xinaris, 1993; Glorfeld, 1995). To confirm the five-factor structure as a final model, we conducted three CFAs, one with the whole sample, one with a randomly split half sample, and one by gender group. We used $RMSEA < .06$, $CFI > .90$, and $TLI > .90$ as model fit indices to confirm the five-factor model.

Multigroup CFA (MG-CFA). Based on the CFA model with five factors that we established using the whole sample, we conducted MI by gender using MG-CFA, which tests several sequences of models, such as the configural invariance model, the metric invariance model, and the scalar invariance model (Meredith, 1993; Thurstone, 1947). As the basis of MI, the model structure for different groups (e.g., gender group) is tested to determine whether the model structure for different groups fits the same way. This is known as testing of the configural invariance model. The configural model is a less restrictive model than are the metric and scalar models. If configural invariance holds the model, metric invariance can be tested to determine whether the factor loadings for different groups are the same. If both configural and metric invariance models are established, the scalar invariance model is followed to determine whether the groups have the same item intercept.

The criteria of RMSEA values between .05 and .08, $CFI > .90$, $TLI > .90$, χ^2 , and ΔCFI were considered acceptable fit indices to

establish MI (Cheung & Rensvold, 2002; Fabrigar et al., 1999; Garver & Mentzer, 1999). In particular, testing scalar invariance is the stage of investigating differential item functioning. Thus, we conducted a differential item functioning (DIF) analysis by calculating logit for DIF measure. The magnitude of the DIF measure of logit was determined by Educational Testing Service (ETS) DIF category: A = negligible, B = slight to moderate ($|DIF| \geq .43$ logits), and C = moderate to large ($|DIF| \geq .64$ logits; Zwirk, Thayer, & Lewis, 1999).

Results

Factor Structures Among High School Students

First, we conducted a descriptive analysis of the BESS data set (i.e., means, standard deviations, skewedness, and kurtosis) for each BESS item. Based on the criteria of skewedness ($|2.0|$) and kurtosis ($|7.0|$; Chou & Bentler, 1995; Curran, West, & Finch, 1996), we found that none of the BESS items showed nonnormality. Then we conducted an initial CFA under study conditions identical to the factor structures from Dowdy et al. (2011), as shown in Figure 2. The model showed a poor fit, $\chi^2(320, N = 4,074) = 8,679.55$, $RMSEA = .08$, $CFI = .84$, $TLI = .82$, which reflected that the factor structure of the high school students' data was different from that of elementary school students' BESS data.

To find an optimal number of factors to represent the high school students' BESS, we performed Horn's (1965) PA and EFA. First, we conducted EFA using all 30 items from two- to five-factor solutions, as shown in Table 1. The five-factor solution was better than the four-factor solution for the high school sample: four

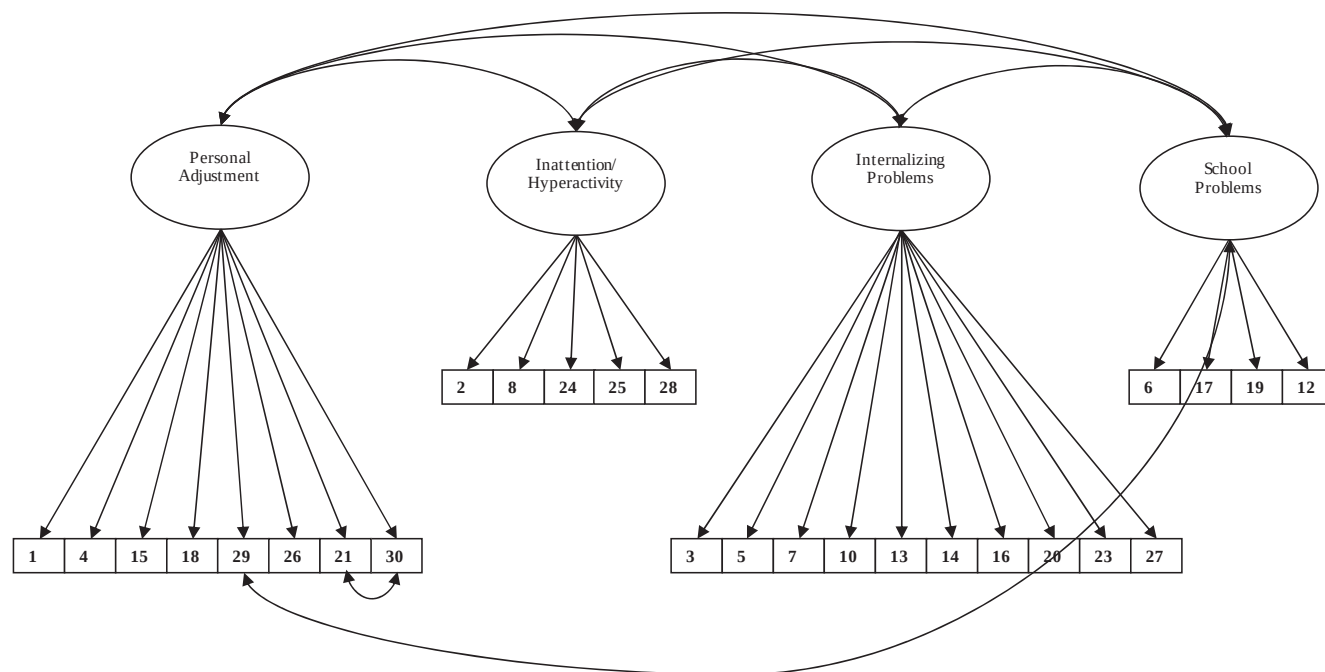


Figure 2. Path diagram of four factors structure of the BESS. From "Factor Structure of the BASC-2 Behavioral and Emotional Screening System Student Form," by E. Dowdy, J. M. Twyford, J. K. Chin, C. A. DiStefano, R. W. Kamphaus, & K. L. Mays, 2011, *Psychological Assessment*, 23, p. 385. Copyright 2011 by the American Psychological Association.

factors, $\chi^2(321, N = 4,074) = 5,692.54$, RMSEA = .06, CFI = .91, TLI = .88, versus five factors, $\chi^2(295, N = 4,074) = 3,655.55$, RMSEA = .05, CFI = .94, TLI = .92, providing appropriate factor loading values ($>.30$) for the items on the five factors (Costello & Osborne, 2005), even though cross loadings were observed for three items (Items 8, 23, and 27; see Table 2). Horn's PA analysis also showed that the five-factor solution was deemed optimal for the high school students' data (see Figure 3). The five factors were personal adjustment, school problems, internalization of problems, inattention or hyperactivity, and relationship with parents. The fifth factor on the five-factor solution comprised three items indicating the relationship with parents: Item 15 ("My parents trust me"), Item 18 ("My parents listen to what I say"), and Item 26 ("My parents are proud of me"). Those three items were included in the personal adjustment factor in the four-factor structure from Dowdy et al.'s (2011) study.

Table 1
Assessment of EFA Model Fit Indices From Two- to Five-Factor Solutions

Solution	χ^2	df	RMSEA	CFI	TLI
2-factor	13,398.260	376	.090	.781	.740
3-factor	8,339.200	348	.081	.862	.831
4-factor	5,692.540	321	.062	.911	.882
5-factor	3,655.550	295	.051	.941	.920

Note. EFA = exploratory factor analysis; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index.

The correlation among the five factors ranged between $-.408$ and $.351$, reflecting that the five factors clustered distinctively (see Table 3). Personal adjustment showed a negative correlation with the factors of inattention or hyperactivity, school problems, and internalization of problems, as would be expected, given that these items reflected positive perceptions of competencies, skills, and attitudes toward self and others.

In addition to measuring the correlations among the factors, we also measured the reliability of items within each factor. First, we tested the assumption of tau equivalence model to estimate Cronbach's alpha as the reliability measure. As seen in Table 4, the chi-square difference in fit between models (null vs. saturated) was statistically significant, indicating that the assumption of the tau equivalence was not met, although other fit indices (i.e., RMSEA, CFI) appeared to be good for the data. This violation of tau equivalence would have resulted in underestimating the reliability of the items if we had used Cronbach's alpha. Thus, we calculated the composite reliability to measure the reliability of items within each factor: .739 for personal adjustment, .813 for school problems, .863 for internalization of problems, .759 for inattention or hyperactivity, and .857 for relationship with parents. The composite reliability has often been used as an alternative measure because its value is slightly higher than that of Cronbach's alpha (R. A. Peterson & Kim, 2013). Composite reliability above .70 is considered adequate (Hair, Anderson, Tatham, & Black, 1998).

In the next step, we ran a CFA using the five factors in the EFA (i.e., personal adjustment, inattention or hyperactivity, school problems, internalization of problems, and relationship with parents); results indicated a poor model fit, $\chi^2(403, N = 4,074) =$

Table 2
Standardized Factor Loadings of Five Factors in EFA

Item synopsis	Personal adjustment	School problems	Internalization of problem	Inattention or hyperactivity	Relationship with parents
1. Good at decision making	.28	-.13	-.05	.02	.19
4. Like looks	.31	.03	-.25	.11	.20
9. Liked by others	.71	-.01	-.05	-.05	-.01
21. Others think I am fun to be with	.77	.02	.03	.03	.02
30. Others respect me	.58	-.03	.01	-.11	.21
6. Interest in quitting school	-.04	.56	.21	.02	-.07
12. School interest	.01	.78	-.09	.06	.00
17. Hate school	-.04	.93	-.01	-.03	.03
19. Teachers unfair	.15	.43	.12	.17	-.03
22. Teachers make me feel stupid	.07	.35	.28	.25	-.02
29. School comfort	.25	-.41	.00	.06	.17
3. Have worries	.02	-.07	.65	.01	-.02
5. Feel out of place	-.27	-.04	.59	-.01	.05
7. Others angry at me	.01	.00	.42	.22	-.10
10. Life getting worse	-.02	.10	.64	-.04	-.20
13. People out to get me	-.08	.05	.50	.21	.00
14. Worry about future	.02	-.09	.67	.02	.01
16. Feel left out	-.23	.01	.52	.13	.04
20. Want to improve but unsuccessful	-.02	.30	.49	.00	.01
23. Blamed for problems out of my control	.02	.11	.44	.33	-.08
27. Am a failure despite effort	.01	.32	.47	-.01	-.01
2. Talk when others talk	.03	.15	-.12	.43	.01
8. Pay attention to teacher	-.05	.33	.11	.42	.02
11. Have trouble sitting still	-.06	-.04	-.01	.77	-.01
24. Am noisy	.03	-.03	.05	.55	.03
25. Get in trouble for inattention	.05	.29	-.03	.51	-.01
28. Have difficulty standing still	-.09	-.02	.03	.70	.01
15. Parents trust me	.00	.00	.02	-.08	.80
18. Parents listen to me	-.03	.05	-.04	-.02	.79
26. Parents are proud of me	.06	-.07	-.01	.01	.81

Note. Bold font indicates factor loading values greater than $|\lambda| \geq .30$ for the items on the five factors structure. Items 8, 23, and 27 represent cross loadings. EFA = exploratory factor analysis.

12,028.42, $p = .00$, RMSEA = .08, CFI = .80, TLI = .80. We reviewed the modification indices and found high expected parameter change (EPC) values among five pairs of items (see Table 5). The EPC statistic represented correlated residuals from improving

a model fit, where a high correlation between items might have been caused by similar item content. For instance, Item 9 asked students whether they can feel whether people like them, and Item 21 asked whether people think they are fun. The students might

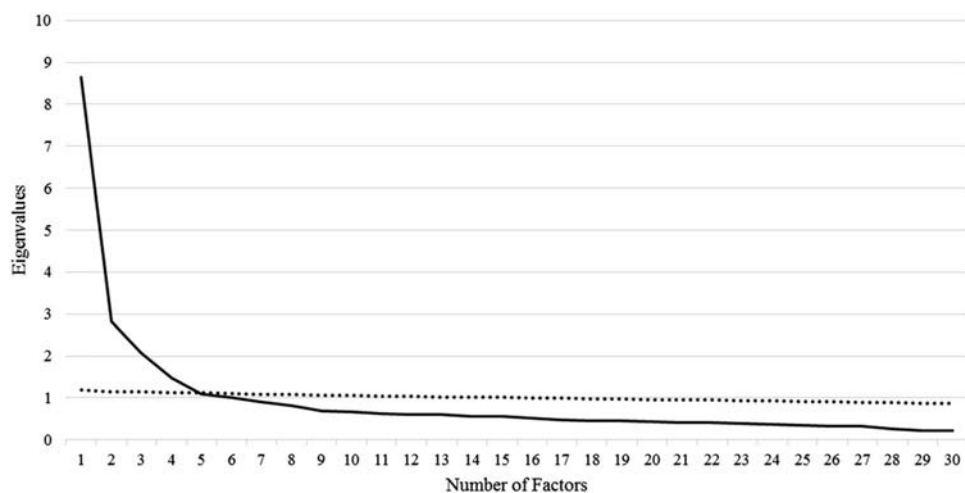


Figure 3. Scree plot of the five-factor solution. This figure illustrates the expected eigenvalues (the dotted line) and the observed eigenvalues (the straight line).

Table 3
Extracted Factor Correlations for the BESS Among High School Students

Factor	1	2	3	4
1. Personal adjustment	—			
2. School problems	-.408	—		
3. Internalization of problem	-.321	.290	—	
4. Inattention or hyperactivity	-.026	.336	.373	—
5. Relationship with parents	.351	-.382	-.408	-.223

Note. BESS = Behavioral and Emotional Screening System.

have confused these two items by thinking that they were being asked the same question, namely, whether “they are liked by others if they are funny.” Therefore, we allowed residual variance correlations of five pairs of items within the five-factor CFA model, and the model fit improved from the original model ($\Delta\chi^2 = 266.92$, $p = .00$, $\Delta CFI = -.10$, $RMSEA = .06$, $CFI = .90$, $TLI = .90$). We followed the cross-validation of the five-factor CFA model under the five pairs of correlated items by using three disaggregated data sets (e.g., random-split sample, male group, and female group). The model fit indices of those three disaggregated data sets also showed acceptable fit (see Table 6).

Tests of Measurement Invariance

Because the BESS high school sample yielded an adequate model fit for a five-factor structure, we conducted MG-CFA to examine MI between male and female adolescents. We used $\Delta\chi^2$, ΔCFI , $RMSEA$, CFI , and TLI to evaluate the model fit of each MI test. Withholding the model under the five pairs of correlated items, we ran the configural model initially with everything free across the gender groups, except for fixing the parameters for factor variances, factor means, and residual variances (Meredith, 1993; Riordan & Vandenberg, 1994; Vandenberg & Lance, 2000). The configural invariance model yielded acceptable fit, $\chi^2(780, N = 4,074) = 6,207.21$, $RMSEA = .06$, $CFI = .91$, $TLI = .92$, indicating that similar factor structures were found for separate gender groups (see Table 7).

Because we established configural invariance as a base model, we assessed metric invariance to compare factor loadings for male and female adolescents. We ran the metric invariance model under

Table 5
Five BESS Item Pairs Showing High Correlations

Pair, item no., and item	Mod. index	StdYX E.P.C
Pair 1 9. I am liked by others. 21. People think I am fun to be with.	202.631	.460
Pair 2 17. I hate school. 12. School is boring.	197.342	.542
Pair 3 14. I worry about what is going to happen. 3. I worry but I don't know why.	146.171	.331
Pair 4 20. I want to do better, but I can't. 27. Even when I try hard, I fail.	190.662	.443
Pair 5 28. I have trouble standing still in lines. 11. I have trouble sitting still.	197.343	.541

Note. BESS = Behavioral and Emotional Screening System; Mod. index = modification index; StdYX = the fully standardized loading; EPC = expected parameter change.

the condition that all factor loadings were constrained equally across gender. All intercepts and residual variances, however, were allowed to vary across gender. The female group (i.e., the reference group) was conditioned constraining the factor variance as 1. The male group was freely estimated. We then examined the invariance of the factor loadings across gender by comparing model fit indices ($\Delta\chi^2$, ΔCFI , $RMSEA$, CFI and TLI) between the configural model and the metric model. The full metric invariance model fit was significantly worse than the configural invariance ($\Delta\chi^2 .81$, $p = .00$, $\Delta CFI = .02$, $RMSEA = .05$, $CFI = .92$, $TLI = .90$). The modification indices suggested that the loading of Item 25 was a source of misfit and should be freed. The second metric invariance model fit with freeing the loading of Item 25 was still significantly poor ($\Delta\chi^2 = 36.87$, $p = .05$, $\Delta CFI = .01$, $RMSEA = .05$, $CFI = .92$, $TLI = .92$). Thus, the factor loading of Item 4 was freed. After freeing the factor loadings of two items consecutively, we found the partial metric invariance (i.e., weak invariance) model fit to be significantly better than the full metric invariance model ($\Delta\chi^2 = 27.42$, $p = .24$, $\Delta CFI = .00$, $RMSEA = .05$, $CFI =$

Table 4
Model Fit Assessment of the Tau Equivalent Model

Model	No. items	No. estimated parameters	χ^2	df	p	CFI	RMSEA estimate	90% CI		DIFF p^a
								Lower limit	Upper limit	
Five-factor (null)	30	135	6,142.403	390	.000	.902	.060	.059	.062	
Tau equivalence (saturated)										
Personal adjustment	5	131	6,197.353	394	.000	.901	.060	.059	.061	.000
School problems	6	130	6,848.712	395	.000	.891	.060	.062	.065	.000
Internalization of problems	10	126	6,485.014	399	.000	.896	.060	.060	.063	.000
Inattention or hyperactivity	6	130	6,849.712	395	.000	.890	.063	.062	.065	.000
Relationship with parents	3	133	6,186.279	392	.000	.901	.060	.059	.062	.000

Note. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; CI = confidence interval; DIF = differential item functioning.

^a The test of difference between null and saturated model.

Table 6
Assessment of Model Fit Indices from Five Factors CFA

Model	χ^2	df	RMSEA	CFI	TLI
Whole	6,142.40	390	.060	.902	.900
Half split	3,026.73	390	.058	.908	.900
Female adolescents	2,771.94	390	.057	.912	.902
Male adolescents	3,441.47	390	.059	.901	.900

Note. CFA = confirmatory factor analysis; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index.

.92, TLI = .92; see Table 7). This evidence showed that items on the BESS Student Form were related to the latent factor equivalently across male and female adolescents, except for two items (Items 25 and 4). Item 25 (“I get into trouble for not paying attention”) was related to the inattention or hyperactivity factor more in the male group than in the female group. Item 4 (“I like the way I look”) was related to the personal adjustment factor more for male adolescents than for female.

In the scalar invariance model, three thresholds from the 4-point range of responses were constrained to be equal. The factor variance and mean for the female group were fixed at 1 and 0, respectively, and the factor variance and mean were estimated for the male group (Widaman & Reise, 1997). All factor loadings and intercepts were constrained equally across gender groups, except for Items 25 and 4, for which factor loadings were freed from the partial metric invariance model. The scalar invariance model fit, however, was significantly worse than the partial metric invariance model ($\Delta\chi^2 = 943.166$, $\Delta\text{CFI} = -.01$, $p = .00$, $\text{RMSEA} = .05$, $\text{CFI} = .91$, $\text{TLI} = .92$). Modification indices suggested that thresholds of 14 items (DIF) should be freed. Based on acceptable model fit indices and no other modifications to free to improve the fit, the scalar invariance model held. However, the scalar invariance model was met partially after freeing the thresholds of 14 items successively ($\Delta\chi^2 = 265.34$, $\Delta\text{CFI} = .01$, $p = .00$, $\text{RMSEA} = .05$, $\text{CFI} = .92$, $\text{TLI} = .92$; see Table 7).

We conducted, based on the findings of the scalar MI model, a follow-up DIF analysis to identify the DIF magnitude of 14 DIF items (see Table 8). According to the DIF contrast and ETS DIF category (Zwick et al., 1999), only five of 14 DIF items showed slight, moderate, or large DIF magnitude: four items (Items 3, 13, 18, and 26) as slight-to-moderate DIF ($|\text{DIF}| \geq 0.43$ logits) and one item (Item 4) as moderate-to-large DIF ($|\text{DIF}| \geq 0.64$ logits). Nine of 14 DIF items showed negligible DIF magnitude (Items 1, 5, 10, 12, 14, 19, 21, 24, and 30). The overall results of the DIF analysis indicated that the female responses produced slightly higher dif-

ficulty in items regarding internalization of problems, and the male responses showed higher difficulty in items regarding personal adjustment. There were only negligible DIF items in the factors of school problems and inattention or hyperactivity.

For the personal adjustment factor, one item (Item 4) showed moderate to large DIF, and three items (Items 1, 21, and 30) showed negligible DIF between female and male adolescents (see Table 8). This result indicated that the item asking about appearance showed higher difficulties for male adolescents: *DIF contrast* = $-.510$ at Threshold 1 (ETS category: B), *DIF contrast* = $-.671$ at Threshold 2 (ETS category: C), and *DIF contrast* = $-.475$ at Threshold 3 (ETS category: B).

For the internalization of problems factor, five items (Items 3, 5, 10, 13, and 14) were identified as negligible, slight, or moderate DIF (see Table 8). Among those five items, Item 3 (“I worry, but I do not know why”) and Item 13 (“I feel like people are out to get me”) showed slight-to-moderate DIF: *DIF contrast* = $.456$ at Threshold 1 (ETS category: B). This finding indicates that items about anxiety and depression showed a higher difficulty level for female adolescents and that items about atypicality showed a higher difficulty level for male adolescents: *DIF category* = $-.442$ at Threshold 3 (ETS category = B). Item 5 (“I feel out of place around people”), Item 10 (“I feel like my life is getting worse and worse”), and Item 14 (“I worry about what is going to happen”) showed negligible DIFs (ETS category: A).

For the relationship with parents factor, two items (Items 18 and 26) showed different difficulty levels between female and male adolescents (see Table 8). Male adolescents had more difficulty responding to Item 18 (“My parents listen to what I say”) than female adolescents did: *DIF category* = $-.225$ at Threshold 2 (ETS category = B). Item 26 (“My parents are proud of me”) indicated that female adolescents had higher difficulty than male adolescents did: *DIF contrast* = $.261$ at Threshold 3 (ETS category = B).

Discussion

As the practice of behavioral and emotional risk screening expands, so, too, does the need to answer foundational questions about the measures in use. In this case, we sought to determine whether the BESS Student Form could be used with some confidence by providing equivalent measurement properties for both female and male adolescents in high school. We began by investigating the factor structure of the BESS Student Form by using the four-factor structure identified previously by Dowdy et al. (2011). Our results were different in that a five-factor structure appeared to be a better fit with the data, which came from a rather large sample

Table 7
Assessment of Measurement Invariance Model Fit Indices using MG-CFA

Model	No. free parameters	χ^2	df	RMSEA	CFI	TLI	$\Delta\chi^2$	ΔCFI
Configural	270	6,207.21	780	.06	.91	.92		
Metric	247	5,281.05	803	.05	.92	.92	27.42	.00
Scalar	183	5,425.59	867	.05	.92	.92	265.34	.01

Note. MG-CFA = multigroup confirmatory factor analysis; RMSEA = root-mean-square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index.

Table 8
DIF Analyses of BESS Items in Scalar Invariance Model

Factor and item	DIF contrast (ETS DIF category) ^a		
	Threshold 1	Threshold 2	Threshold 3
Personal adjustment			
1. I am good at making decisions.		-.362 (A)	-.151 (A)
4. I like the way I look.	-.510 (B)	-.671 (C)	-.475 (B)
21. People think I am fun to be with.	-.184 (A)	-.184 (A)	.085 (A)
30. Others have respect for me.			.330 (A)
School problems			
12. School is boring.		-.26 (A)	-.218 (A)
19. Teachers are unfair.			-.267 (A)
Internalization of problem			
3. I worry, but I do not know why.	.456 (B)	.352 (A)	.378 (A)
5. I feel out of place around people.	.233 (A)		
10. I feel like my life is getting worse and worse.	.278 (A)		
13. I feel like people are out to get me.		-.400 (A)	-.442 (B)
14. I worry about what is going to happen.	.292 (A)	.280 (A)	
Inattention or hyperactivity			
24. People tell me that I am too noisy.	-.140 (A)	-.147 (A)	-.147 (A)
Relationship with parents			
18. My parents listen to what I say.		-.225 (B)	
26. My parents are proud of me.			.261 (B)

Note. DIF = differential item functioning; BESS = Behavioral and Emotional Screening System; ETS = Educational Testing Service.

^a From "An Empirical Bayes Approach to Mantel Haenszel DIF Analysis," by R. Zwirk, D. T. Thayer, & C. Lewis, 1999, *Journal of Educational Measurement*, 36, p. 1–28. Copyright 1999 by John Wiley & Sons, Inc.

of more than 4,000 students from four high schools in the LAUSD. Based on item content, we labeled the fifth factor as relationship with parents. The same items were subsumed in the Dowdy et al. study under the personal adjustment factor, which is a measure of prosocial and interpersonal competencies (Reynolds & Kamphaus, 2004). The difference in factor structures could be attributed to typical developmental changes in parent and adolescent relationships as students enter adolescence (Steinberg, 2001).

Overall, we conclude that the current version of the BESS Student Form is a partially invariant screening tool, regardless of gender. We use the term *partial* to denote the fact that the measure could be improved based on the finding that five of the 30 items functioned somewhat differently for female and male adolescents. We think that further examination of these five DIF items might provide insight into how the female and male adolescents' responses were differentiated on items regarding their relationships with their parents, self-esteem, self-reliance, depression or anxiety, and atypicality. By way of example, an item asking about satisfaction with one's appearance produced statistically significantly different item difficulty levels between gender groups, indicating that the threshold estimates of the male responses were higher than the female. This finding does not imply that male adolescents rated themselves more favorably than female adolescents did on their looks (i.e., perception of the physical attractiveness). Instead, the item about appearance had higher difficulty for male adolescents, even if female self-ratings of this item were on average lower ($M = 2.71$) than male ($M = 3.13$). The other items found to demonstrate DIF in the scalar analyses showed that female adolescents also had more problems on items related with anxiety and depression.

This study has two primary implications. First, we found no compelling evidence that the BESS Student Form is unusable at

this time, particularly because the need for screening is great. Schools in the United States in particular are eager to engage in any practice that might improve school safety and security for children, which are two primary goals of behavioral and emotional risk screening (O'Connell, Boat, & Warner, 2009). Our second implication has to do with test development and revisions. Items demonstrating DIF in this study should be prioritized for further evaluation as to whether they should be included in revisions. Some pairs of items showing content similarity, even though those items have slightly different content, should also be evaluated for revisions to make the instrument usable with greater confidence for screening programs.

The primary limitation of this study was the composition of the sample, described by school records as predominantly Hispanic and Latino/a American, thus limiting the generalizability of the results to other population subgroups. The composition of the sample served as a potential confound in that ethnicity, not gender differences per se, could ultimately be the cause of the observed DIF. Nonetheless, our findings demonstrate the importance of measurement invariance testing by revealing partial metric and scalar invariance by gender group.

This study brings attention to the need for the routine measurement invariance evaluation of behavioral and emotional risk screening measures such as the BESS Student Form. Further, the level of measurement invariance (e.g., effect size) is also of interest to examine how much the extent of observed MI matters. The need for additional measurement invariance studies with different target subpopulations identified by grade level; geographic location; ethnicity; or differences attributable to student, teacher, or parent informants is great. In particular, analyzing the perspectives of the three informant groups (e.g., students, parents, and teachers) should be a high priority because a large body of research has

produced evidence that teachers and parents have different perspectives based on the context in which they interact with children (Lawson, 2003; Petrakos & Lehrer, 2011).

References

- Atkins, M. S., Frazier, S. L., Adil, J. A., & Talbott, E. (2003). School-based mental health services in urban communities. In M. D. Weist, S. W. Evans, & N. A. Lever (Eds.), *Handbook of school mental health: Advancing practice and research: Issues in clinical child psychology* (pp. 165–178). New York, NY: Kluwer Academic/Plenum Press.
- 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. <http://dx.doi.org/10.1007/s10864-007-9058-6>
- Catalano, R. F., Haggerty, K. P., Oesterle, S., Fleming, C. B., & Hawkins, J. D. (2004). The importance of bonding to school for healthy development: Findings from the Social Development Research Group. *Journal of School Health*, 74, 252–261. <http://dx.doi.org/10.1111/j.1746-1561.2004.tb08281.x>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness of fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5
- Chou, C., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–55). Thousand Oaks, CA: Sage.
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10, 7. Retrieved from <http://pareonline.net/pdf/v10n7a.pdf>
- Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement*, 53, 585–596. <http://dx.doi.org/10.1177/0013164493053003001>
- 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. <http://dx.doi.org/10.1037/1082-989X.1.1.16>
- Dever, B. V., Mays, K. L., Kamphaus, R. W., & Dowdy, E. (2012). The factor structure of the BASC-2 Behavioral and Emotional Screening System Teacher Form, Child/Adolescent. *Journal of Psychoeducational Assessment*, 30, 488–495. <http://dx.doi.org/10.1177/0734282912438869>
- Dowdy, E., Twyford, J. M., Chin, J. K., DiStefano, C. A., Kamphaus, R. W., & Mays, K. L. (2011). Factor structure of the BASC-2 Behavioral and Emotional Screening System Student Form. *Psychological Assessment*, 23, 379–387. <http://dx.doi.org/10.1037/a0021843>
- Edelbrock, C. S., & Achenbach, T. M. (1980). A typology of child behavior profile patterns: Distribution and correlates for distributed children aged six–sixteen. *Journal of Abnormal Psychology*, 8, 441–470.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272–299. <http://dx.doi.org/10.1037/1082-989X.4.3.272>
- Frick, P. J., Burns, C., & Kamphaus, R. W. (2009). *Clinical assessment of child and adolescent personality and behavior* (2nd ed.). New York, NY: Springer.
- Garver, M. S., & Mentzer, J. T. (1999). Logistics research methods: Employing structural equation modeling to test for construct validity. *Journal of Business Logistics*, 20, 33–57.
- Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55, 377–393. <http://dx.doi.org/10.1177/0013164495055003002>
- Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.
- Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis*. Englewood Cliffs, NJ: Prentice Hall.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185. <http://dx.doi.org/10.1007/BF02289447>
- Hyde, J. S., Mezulis, A. H., & Abramson, L. Y. (2008). The ABCs of depression: Integrating affective, biological, and cognitive models to explain the emergence of the gender difference in depression. *Psychological Review*, 115, 291–313. <http://dx.doi.org/10.1037/0033-295X.115.2.291>
- Kamphaus, R. W., & Reynolds, C. R. (2007). *BASC-2 Behavioral and Emotional Screening System*. Minneapolis, MN: Pearson Assessment.
- Lane, K. L., Carter, E. W., Pierson, M. R., & Glaeser, B. C. (2006). Academic, social, and behavioral characteristics of high school students with emotional disturbances or learning disabilities. *Journal of Emotional and Behavioral Disorders*, 14, 108–117. <http://dx.doi.org/10.1177/10634266060140020101>
- Lawson, M. A. (2003). School-family relations in context: Parents and teacher perceptions of parent involvement. *Urban Education*, 38, 77–133. <http://dx.doi.org/10.1177/0042085902238687>
- McIntosh, K., Flannery, K., Sugai, G., Braun, D., & Cochrane, K. (2008). Relationships between academics and problem behavior in the transition from middle school to high school. *Journal of Positive Behavior Interventions*, 10, 243–255. <http://dx.doi.org/10.1177/1098300708318961>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543. <http://dx.doi.org/10.1007/BF02294825>
- Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9, 93–115. <http://dx.doi.org/10.1037/1082-989X.9.1.93>
- 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 Academy Press.
- Peterson, D. R. (1961). Behavior problems of middle childhood. *Journal of Consulting Psychology*, 25, 205–209. <http://dx.doi.org/10.1037/h0038994>
- Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology*, 98, 194–198. <http://dx.doi.org/10.1037/a0030767>
- Petrakos, H. H., & Lehrer, J. S. (2011). Parents' and teachers' perceptions of transition practices in kindergarten. *Exceptionality Education International*, 21, 62–73. Retrieved from <http://ir.lib.uwo.ca/eei/vol21/iss2/7>
- Plecki, M. L., Elfers, A. M., & Nakamura, Y. (2012). Using evidence for teacher education program improvement and accountability: An illustrative case of the role of value-added measures. *Journal of Teacher Education*, 63, 318–334. <http://dx.doi.org/10.1177/0022487112447110>
- Quay, H. C. (1987). Patterns of delinquent behavior. In H. C. Quay (Ed.), *Handbook of juvenile delinquency* (pp. 118–138). New York, NY: Wiley.
- Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children—2nd ed. (BASC-2)*. Bloomington, MN: Pearson.
- Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, 20, 643–671. <http://dx.doi.org/10.1177/014920639402000307>
- Steinberg, L. (2001). We know some things: Parent-adolescent relationships in retrospect and prospect. *Journal of Research on Adolescence*, 11, 1–19. <http://dx.doi.org/10.1111/1532-7795.00001>

- Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago, IL: University of Chicago Press.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–70. <http://dx.doi.org/10.1177/109442810031002>
- Wagner, M., Kutash, K., Duchnowski, A., & Epstein, M. (2005). The special education elementary longitudinal study and the national longitudinal transition study: Study designs and implications for children and youth with emotional disturbance. *Journal of Emotional and Behavioral Disorders*, 13, 25–41. <http://dx.doi.org/10.1177/10634266050130010301>
- Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). <http://dx.doi.org/10.1037/10222-009>
- Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel Haenszel DIF analysis. *Journal of Educational Measurement*, 36, 1–28. <http://dx.doi.org/10.1111/j.1745-3984.1999.tb00543.x>

Received April 26, 2016

Revision received November 8, 2016

Accepted February 7, 2017 ■