7-2016

Factor Analysis of the Preschool Behavioral and Emotional Rating Scale for Children in Head Start Programs

Cynthia J. Cress
University of Nebraska-Lincoln, ccress1@unl.edu

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

Michael Epstein
University of Nebraska - Lincoln, mepstein1@unl.edu

Published in *Journal of Psychoeducational Assessment* 34:5 (2016), pp 473–486.
doi 10.1177/0734282915617630

Factor Analysis of the Preschool Behavioral and Emotional Rating Scale for Children in Head Start Programs

Cynthia Cress, Matthew C. Lambert, and Michael H. Epstein

University of Nebraska–Lincoln, USA

Corresponding author — Cynthia Cress, Special Education and Communication Disorders, University of Nebraska–Lincoln, 202B Barkley Memorial Center, Lincoln, NE 68583-0738, USA; email ccress1@unl.edu

Abstract
Strength-based assessment of behaviors in preschool children provides evidence of emotional and behavioral skills in children, rather than focusing primarily on weaknesses identified by deficit-based assessments. The Preschool Behavioral and Emotional Rating Scales (PreBERS) is a normative assessment of emotional and behavioral strengths in preschool children. The PreBERS has well-established reliability and validity for typically developing children as well as children with identified special education needs, but this has not yet been established for children in Head Start programs, who tend to be at high risk for development of emotional and behavioral concerns. This study explores the factorial validity of the PreBERS scores for a large sample of children participating in Head Start programs around the United States. Results not only confirm the fit of the four-factor model of the PreBERS for this population, but also demonstrate the application of a bifactor model to the structure of the PreBERS which, in turn, allows for the computation of model-based reliability estimates for the four subscales (Emotional Regulation, School Readiness, Social Confidence, Family Involvement) and overall strength index score. The implications suggest that the PreBERS items are reliable scores that can be used to identify behavioral strengths in preschool children in Head Start, and support planning of interventions to selectively address component skills to promote child social and academic success.

Keywords: preschool children, behavioral strengths, strength-based assessment, psychometrics, factor analysis

The education of young children and the emergence of early childhood education have recently taken on significant importance. Nationwide local, state, and federal programs as well as foundations have developed programs designed to improve the school preparedness of all children. These programs are designed to enhance the health, language, academic, behavior, and overall functioning of children.

An important part of early childhood programs is assessment which is defined as gathering information to make informed instructional decisions about children (Salvia, Ysseldyke, & Bolt, 2013). These decisions involve screening, diagnosis, identification, intervention planning, and evaluation.

The assessment of social, emotional, and behavioral functioning is particularly important at the preschool level because of the number of children impacted and the persistence and stability of behavior problems (McCain, Mustard, & Shanker, 2007). Prevalence studies have indicated that upward of 15% of preschool children demonstrate some type of emotional or behavioral problem (Powell, Dunlap, & Fox, 2006). In addition, studies of children in Head Start programs indicate upward of 15% to 30% present some type of emotional or behavioral problem (e.g., Qi & Kaiser, 2003). Moreover, investigators have found that behavior problems are quite persistent over time, namely, that children who evidence behavior problems in preschool settings are very likely to demonstrate behavior problems throughout school (Carter, Briggs-Gowan, & Davis, 2004). Moreover, children who present behavior problems are also very likely to present with academic, family, mental health, and physical health problems (Kauffman & Landrum, 2009).

In light of the prevalence of behavior problems among young children and the poor prognosis for children who evidence these problems, several investigators have developed behavior rating scales to screen, identify, and diagnose children. Most of these instruments rely on a deficit model for assessment, which may limit the type of information collected about the child including not identifying what the child does well. Indeed, identifying the positive aspects or personal strengths of the child could be very helpful in designing and evaluating interventions and services.

A number of years ago the Working Group on Developmental Assessment set forth basic principles for assessing young children and identified one core principle devoted to strength-based assessment (Greenspan & Meisels, 1996). The group stated, "The assessment process should identify the child's current competencies and strengths, as well as the competencies that will constitute developmental progression in a continuous growth model of development" (Meisels & Atkins-Burnett, 2000, p. 232). The National Association for the Education of Young Children (NAEYC) and the National Association of Early Childhood Specialists in State Departments of Education have articulated a similar position. Strength-based assessment adds to the intervention planning process in several ways including focusing efforts on a child's strengths as opposed to weaknesses and pointing the direction for growth, enhancing the motivation and engagement of parents and caregivers to the intervention process, and identifying areas for progress monitoring of child behavior during interventions (Epstein, 2004).

In response to the need for strength-based measures to assess the behavioral functioning of preschoolers, several investigators developed such assessments. These include the Child Preference Indicators (Moss, 2006), Devereux Early Childhood Assessment Program (LeBuffe & Naglieri, 1998), Infant and Toddler Social and Emotional Assessment (Briggs-Gowan & Carter, 1998), and Vineland Social-Emotional Early Childhood Scales (Sparrow, Balla, & Cicchetti, 1998). In addition, a few measures such as the Child Behavior Checklist (Achenbach & Rescorla, 2001) and the Strengths and Difficulties Questionnaire (Goodman, 1997) include both strength and deficit oriented items. Perhaps one of the most recent and widely used strength-based assessment for young children is the Preschool Behavioral and Emotional Rating Scale (PreBERS; Epstein & Synhorst, 2009), a standardized and norm-referenced assessment of emotional and behavioral strengths of preschool children. The PreBERS stands out from these other strength-based assessments in its specific focus on multiple emotional and behavioral strengths in preschoolers, as well as its consistency in types of strengths assessed with a school-aged behavioral strength assessment, the Behavioral and Emotional Rating Scale (Epstein, 2004).

The psychometric characteristics of the item responses and scores were established in a series of investigations by the test developers. First, the developers followed a multi-step process to prepare the 42-items including a review of the available strength-based measures, a comprehensive review of the research on the development of social-emotional behavior in preschoolers, a qualitative study of the views of preschool teachers and parents, an item analysis of a prototype scale, and an exploratory factor analysis (Epstein, Synhorst, Cress, & Allen, 2009). Second, reliability studies found the internal consistency (subscale range .84 to .98; Epstein et al., 2009), test-retest reliability (subscale range .80 to .89), and inter-rater reliability (subscale range .71 to .85) of the PreBERS scores to be acceptable (Cress, Epstein, & Synhorst, 2010). Third, convergent validity was assessed with the Child Behavior Checklist Caregiver–Teacher Report Form (C-TRF; Achenbach & Rescorla, 2000) with moderate to large correlations from parent (i.e., −.37 and −.78; Nordness, Epstein, & Synhorst, 2009) and teacher raters (−.61 and −.84; Epstein & Synhorst, 2009).

Previous research has supported the psychometric functioning of the PreBERS scores. These studies were conducted first on children in typical preschools (Epstein et al., 2009) and then with children in Head Start classrooms (Griffith et al., 2010). However, in these earlier studies, the factor models (i.e., measurement models) were based on the relationships between subscale standard scores as opposed to item-level responses, and only tested the hypothesis that the four subscale scores were correlated because of a general strength index factor (see Epstein & Synhorst, 2009). The models did not account for error variances at the item level and the inter-correlation of these variances (Brown, 2006), and therefore did not test the hypothesis about whether the items formed four distinct, but correlated factors. A more exacting analysis is to model the internal structure using item-level responses. The importance of establishing a tenable measurement model for the PreBERS scores cannot be overstated since the measurement model serves as the "basis and rationale for arriving at the composite [scores]" (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999, p. 20) and is a prerequisite for assessing score reliability (Slaney & Maraun, 2008). Therefore, in the present study we analyzed the internal structure (i.e., factorial validity) of the instrument using item-level responses in a confirmatory factor analysis (CFA) framework.

Method

Participants

The participants were 909 children ranging in age between 3 years 0 months to 5 years 11 months from Head Start programs nationwide. Data were collected between 2006 and 2007 from several states including California, Indiana, Kansas, Kentucky, Missouri, Montana, Nebraska, New Mexico, Oklahoma, Pennsylvania, South Dakota, Tennessee, Utah, Vermont, Wisconsin, and Wyoming. The sample was almost evenly divided between males (51%) and females (49%), with a mean age of 3.93 years (SD = 0.74). The racial and ethnic composition of the sample was 31.7% European American, 26.1% African American, 1.2% Asian, 8.7% multi-racial, and 32.3% from other racial backgrounds including individuals of Native American decent. In addition, 32.4% of children were identified as Hispanic. Just over 16% of children were identified with a disability; the most prevalent disability was speech and language disorders (10.3% of participants) followed by developmental delay (4.6%) and then attention deficit disorders (1.4%) and behavioral and emotional disorders (1.4%).

The sampling strategy resulted in a national sample of children in Head Start programs. The profiles of these children were similar to other children in Head Start programs (O'Brien et al., 2002) with respect to gender, race, ethnicity, and disability status. However, the sample was somewhat under-representative of the northeast and over-representative of the south with slightly more 5-year-olds than the Head Start population.

Procedures

Before data collection commenced, the university institutional review board (IRB) approved the child recruitment procedures. A national recruitment strategy to obtain data from Head Start centers across the United States was implemented. Specifically, directors of local Head Start programs throughout the United States were emailed and asked to participate. For directors who failed to respond in 2 weeks, a follow-up email was sent. Head Start directors from 16 states agreed to assist in data collection. Letters were sent to these individuals indicating how to serve as site administrators, how to recruit staff to participate, and how to train staff to complete the PreBERS rating form. Teachers were asked to read and sign consent forms before they started their participation, and teachers only rated students for whom they were the primary Head Start instructor, to avoid duplication of student rankings by multiple teachers.

Head Start teachers were given further instruction on how to identify the children from their classroom. Specifically, the teachers were provided the following instructions:

> First, decide how many students you wish to rate. Then, start either at the top or bottom of your class roster and rate every other child. Do not skip any child unless you have known this child less than 2 months. Stop selecting and rating children when you have reached the number of children you wished to rate.

The selection process was implemented to reduce the likelihood of selection bias of the teachers who were providing the ratings. Teachers completed paper-and-pencil forms and returned these forms by mail to the experimenters. The instruction for teachers to select the numbers of children to rate was implemented to increase teacher participation in the research activity, and avoid the likelihood of teacher refusal of the task if the suggested target number was considered too time-consuming for that teacher. For the same reason, teachers were given options for returning forms anonymously, and no data were collected from the teachers, not even teacher ID numbers. Therefore, exact numbers of teachers completing rankings or numbers of children rated per teacher could not be determined. In sum, the process used to secure the participation of the Head Start directors and teachers resulted in our securing a convenience sample.

Instrument

The PreBERS is a standardized, norm-referenced assessment instrument developed to measure the emotional and behavioral strengths of preschool children. The PreBERS measures the following four emotional-behavioral strengths in preschool children: Emotional Regulation, School Readiness, Social Confidence, and Family Involvement (Epstein & Synhorst, 2009). The instrument has 42 items that are judged by preschool teachers and other adults who are familiar with the child. Raters complete the 42-item instrument in about 10 min. The PreBERS measures four functions: 13 items of Emotional Regulation (e.g., "shows concern for the feelings of others"); 13 items of School Readiness (e.g., "works independently"); 9 items of Social Confidence (e.g., "asks others to play"); and 7 items of Family Involvement (e.g., "interacts positively with parents"). Raters judge the items on a 4-point Likert-type scale (0 = *not at all like the child*; 1 = *not much like the child*; 2 = *like the child*; 3 = *very much like the child*). A standard score (M = 10) and a standard deviation (SD = 3) are determined for each subscale as well as a mean (100) and a standard deviation (15) for the total strength index.

Analysis Plan

Mplus v7.11 (Muthen & Muthen, 1998-2014) was used to investigate the reliability and factorial validity of the PreBERS scores using a CFA approach. CFA modeling yields a comprehensive evaluation of the internal structure of the assessment (i.e., factorial validity), establishes a tenable measurement model, and provides estimates used to calculate omega (ω) and omega hierarchical (ω_h) reliability coefficients (McDonald, 1978, 1999; Zinbarg, Yovel, Revelle, & McDonald, 2006).

CFA. The focus of this study was to examine the hypothesized correlated-factor model with four latent factors representing each subscale on the PreBERS. However, some methodologists suggest that comparing a "target" hypothesized model with alternative models allows researchers to explore multiple *plausible* latent structures of the assessment data which may assist with interpretation of the target model (Marsh, Hau, & Grayson, 2005). So as a basis for comparison, two alternative factor models were fit to the data: (a) a single-factor model and (b) a bifactor model (i.e., nested-factor model; general-specific model) with a general strength factor and four group factors which were all constrained to be orthogonal (i.e., uncorrelated).

As a basis for comparison, the single-factor model was used to test whether the most general construct of behavioral and emotional strengths was consistent with the item response variances. A bifactor version of the correlated-factor model was fit to provide an additional perspective on the internal structure of the assessment by partitioning item response variance into common sources (Reise, 2012). The bifactor model assumes that each item response has two sources of variance: the general factor and one of the four group factors (Emotional Regulation, School Readiness, etc.). Thus, the bifactor model can complement a correlated factors model "by evaluating whether item response variance is due to a general construct versus group factors" (Brouwer, Meijer, & Zevalkink, 2013, p. 138).

Since items were measured on a 4-point response scale, we treated the ratings as ordinal rather than continuous indicators of the latent factors. Accordingly, we used weighted least squares with mean and variance adjustments (WLSMV) to estimate the models, and the factors were scaled using a fixed mean and variance approach. All models were specified without correlated residual variances between items. Missing data were minimal (<1%) and excluded from the analysis using a pairwise-present method as is default in Mplus when using the WLSMV estimator.

Chi-square (χ^2), the comparative fit index (CFI; Bentler, 1990), the Tucker–Lewis index (TLI; Tucker & Lewis, 1973), and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) were used to assess model fit. Chi-square represents an exact test of fit and a non-significant value indicates that the model fits the data acceptably; however, χ^2 is typically regarded as too conservative in applied research when sample sizes are large (Browne & Cudeck, 1993). CFI and TLI are comparative fit indices representing the degree of improvement over the worst fitting model (Boomsma, 2000). Both indices are scaled from 0 to 1 with values closer to 1 indicating better fit. An *acceptable* fitting model has a CFI and TLI greater than or equal to 0.90 (Browne & Cudeck, 1993), a *close* fitting model has a value greater than or equal to 0.95. RMSEA represents the degree of model misfit and is reported on a scale of 0 to 1; values closer to zero indicate better fit. Values less than 0.08 are considered *acceptable* (Hu & Bentler, 1999). The 90% confidence interval for the RMSEA was also computed; the upper bound of the interval should be less than 0.08 for acceptable fit. The DIFFTEST chi-square difference test ($\Delta\chi^2$; Muthen & Muthen, 1998-2014) was computed to evaluate the fit of nested models (e.g., the one-factor vs. the four-factor model or the bifactor vs. the four-factor model). Non-significant tests indicate that the fit of the two models being compared do not differ statistically. Statistically significant tests indicate that one model fits the data more closely than the other model.

Model-based reliability. Omega and omega hierarchical reliability coefficients indicate the degree to which the scale scores (i.e., sum scores) precisely measure the target constructs (i.e., the proportion of the sum score variance that can be attributed to the target construct [true score]). Cronbach's alpha has been reported for this population in prior studies of the PreBERS (Griffith et al., 2010), so in this study we computed CFA-based ω estimates to compensate for the known limitations of Cronbach's alpha as an estimate of reliability under conditions where factor loadings are unequal across items or residual variances are correlated (Green & Hershberger, 2000; Raykov, 1997; Sijtsma, 2009).

Table 1. CFA Model Fit Indicators.

	χ^2 (*df*)	$\Delta\chi^2$	CFI	TLI	RMSEA [90% CI]
One-factor	9,718.66 (819)	—	0.892	0.886	0.109 [.107, .111]
Four-factor	6,008.51 (813)	936.12*	0.937	0.933	0.084 [.082, .086]
Bifactor	4,973.16 (777)	927.97*	0.949	0.943	0.077 [.075, .079]

$\Delta\chi 2$ was calculated as 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. The bifactor model was compared with the four-factor model. CFA = confirmatory factor analysis; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval.
* $p < .001$

Omega and omega hierarchical were based on the bifactor CFA model and calculated using Equations 1 and 2, respectively, where, $\lambda_{i\mathrm{GEN}}$ is the loading for each item on the general factor, $\lambda_{i\mathrm{GRP}}$ is the loading for each item on each group factor, and θ_i^2 is the error variance for each item. As the equations indicate, the ω coefficient (Equation 1) includes all sources of common variance in the numerator whereas the ω_h coefficient (Equation 2) only includes a single variance component in the numerator. Note that estimating ω for a subscale score includes the factor loadings for the one group factor *and* the general factor (but not the other group factors; that is, ω for a subscale is influenced by the general factor as well as the one group factor variance). See McDonald (1999), Brunner, Nagy, and Wilhelm (2012), or Reise (2012) for a more through explanation of calculating and interpreting ω reliability estimates using parameters from CFA models.

$$\omega = \frac{\left(\sum \lambda_{i\mathrm{GEN}}\right)^2 + \left(\sum \lambda_{i\mathrm{GRP}_1}\right)^2 \ldots + \left(\sum \lambda_{i\mathrm{GRP}_p}\right)^2}{\left(\sum \lambda_{i\mathrm{GEN}}\right)^2 + \left(\sum \lambda_{i\mathrm{GRP}_1}\right)^2 \ldots + \left(\sum \lambda_{i\mathrm{GRP}_p}\right)^2 + \sum \theta_i^2} \tag{1}$$

$$\omega_h = \frac{\left(\sum \lambda_{i\mathrm{GEN}}\right)^2}{\left(\sum \lambda_{i\mathrm{GEN}}\right)^2 + \left(\sum \lambda_{i\mathrm{GRP}_1}\right)^2 \ldots + \left(\sum \lambda_{i\mathrm{GRP}_p}\right)^2 + \sum \theta_i^2} \tag{2}$$

Results

Table 1 and Table 2 contain the CFA model fit indicators and the factor loadings for each model, respectively. Specifically, Table 1 lists the chi-square test of model fit, chi-square difference test, CFI, TLI, and the RMSEA. Table 2 lists the factor loadings for the single-factor model, the four-factor model, and both general and group factor loadings for the bifactor model.

One-Factor Model

The most constrained model, the one-factor model, did not fit the data acceptably as indicated by the CFI and TLI values less than 0.90 and the RMSEA value greater than 0.08. Interestingly, all of the factor loadings are large (>.50) to very large (>.85), seemingly indicating a strong single factor. Even with strong factor loadings, model misfit was caused by the substantially correlated residual variances for the items on the same subscales. The correlated residuals suggest multidimensionality due to the content heterogeneity of the assessment items.

Table 2. Factor Loadings by CFA Model.

	One-factor	Four-Factor	Bifactor	
			General	Group
Emotional Regulation				
Controls anger	.72	.75	.62	.48
Expresses remorse	.77	.80	.73	.31
Shows concern	.85	.89	.82	.31
Reacts calmly	.80	.83	.67	.55
Handles frustration	.78	.82	.74	.33
Takes turns	.81	.85	.75	.37
Accepts responsibility	.85	.89	.80	.35
Loses gracefully	.81	.84	.70	.53
Accepts "no"	.82	.85	.71	.52
Respects others	.88	.92	.79	.47
Shares	.87	.89	.76	.50
Apologizes	.80	.84	.79	.21
Is kind	.84	.88	.77	.41
School Readiness				
Understands words	.75	.79	.70	.43
Converses	.80	.84	.76	.40
Persists with tasks	.76	.82	.79	.07
Hygiene skills	.65	.70	.67	.09
Understands sentences	.76	.80	.69	.52
Listens to others	.76	.81	.75	.30
Pays attention	.83	.88	.86	.10
Listens to stories	.80	.84	.82	.10
Follows directions	.83	.88	.83	.27
Retells stories	.82	.86	.76	.45
Uses details	.83	.86	.75	.53
Works independently	.67	.71	.69	.05
Uses numbers/colors	.69	.73	.64	.46
Social Confidence				
Is self-confident	.71	.77	.70	.52
Acknowledges feelings	.65	.71	.66	.24
Asks for help	.64	.70	.65	.24
Stands up for self	.54	.59	.51	.56
Accepts closeness	.66	.72	.68	.14
Identifies feelings	.78	.85	.80	.17
Makes friends	.80	.87	.83	.14
Asks other to play	.77	.84	.79	.21
Enthusiastic about life	.72	.78	.73	.29
Family Involvement				
Sense of belonging	.78	.86	.63	.69
Trusts sign. person	.77	.85	.62	.68
Positive relationships	.78	.88	.72	.54
Positive with parents	.74	.84	.69	.49
Involved in discussions	.74	.87	.75	.24
Positive with siblings	.68	.79	.70	.12
Participates in activities	.77	.88	.74	.44

CFA = confirmatory factor analysis

Four-Factor Model

The four-factor model fit the data acceptably in terms of CFI and TLI indices (>0.90) and represented a meaningful improvement over the one-factor model as indicated by the statistically significant chi-square difference test, $\Delta\chi^2_{(6)} = 936.12$, $p < .0001$. The factor loadings for the four-factor model were all large (>0.50) and positive suggesting that the items are good indicators of the factors. However, given the relatively large RMSEA value, the adequacy of this model is questionable; as the 90% confidence interval indicates, if we were to resample from this population repeatedly, nearly all of the samples would demonstrate an unacceptably high RMSEA. The four factors were highly correlated ranging from $r = .73$ (School Readiness with Family Involvement) to $r = .87$ (School Readiness with Social Confidence) indicating that the factors share a substantial amount of common variance suggesting that a hierarchical factor structure might better represent the assessment data.

Bifactor Model

The last model, the bifactor version of the four-factor model, is the best fitting of the models tested as indicated by the acceptable CFI and TLI, and the large difference in fit between the bifactor and the four-factor model, $\Delta\chi^2_{(36)} = 927.97$, $p < .0001$. The bifactor model also had an acceptable RMSEA as expressed by the point estimate and the confidence interval. Examining the factor loadings for the bifactor model reveals that a strong general factor and weaker group factors merged (see Table 2). Overall, 78% of the explained common variance (ECV) was attributed to the General Strength factor, 8% was attributed to the Emotional Regulation factor, 6% to the Family Involvement factor, 5% to the School Readiness factor, and 3% to the Social Confidence factor. Moreover, all factor loadings for the general factor were large (range = .47–.85; *Mdn* = .70) and most loadings for the group factors were moderate to large (range = .06–.71; *Mdn* = .41). However, a few items demonstrated high general factor loadings, but low group factor loadings that were not statistically significant indicating that, after controlling for the general factor, the residual variance of the item is unrelated to the group factor. Examples of items that do not load significantly onto the group factors include *Persists With Tasks* (School Readiness), *Works Independently* (School Readiness), and *Accepts Closeness* (Social Confidence).

Item-parameter invariance. An important assumption underlying the validity of the bifactor model relates to *item-parameter invariance* for the general factor (Reise, 2012), which represents the degree to which the general factor measures the same construct when only a subset of the assessment items are used. It has been suggested that item-parameter invariance indicates that ratings are "*unidimensional enough* so that the item parameter estimates properly reflect the latent trait held in common among the items and are not biased by additional common dimensions caused by clusters of items with similar content" (Reise, Horan, & Blanchard, 2010, p. 216). A lack of invariance tends to indicate meaningful multi-dimensionality in the assessment data. Prior to interpreting the structural parameters of the bifactor model, item-parameter invariance was evaluated for several subsets of items (Reise, 2012). The general factor loadings for each subset are presented in Table 3. Overall, the factor loadings are quite consistent across the subsets, but the variation that is present in the loadings seems to suggest that the general factor loadings are not strictly invariant. That is to say, the general factor has a slightly different meaning depending on which of the group factors are included in the measurement model.

Reliability estimates. Given an acceptable fitting bifactor model and the reasonableness of item-parameter invariance, ω and ω_h were computed based on this model (see Table 4). Omega estimates, which indicate the proportion of scale score variance that can be attributed to true score

Table 3. Invariance Analysis of General Factor Loadings.

	Full model	Alternative Model 1	Alternative Model 2	Alternative Model 3	Alternative Model 4
Emotional Regulation					
Controls anger	.62	.60	.66	—	.71
Expresses remorse	.73	.73	.82	—	.82
Shows concern	.82	.81	.90	—	.90
Reacts calmly	.67	.65	.67	—	.75
Handles frustration	.74	.74	.70	—	.78
Takes turns	.75	.75	.74	—	.81
Accepts responsibility	.80	.79	.81	—	.86
Loses gracefully	.70	.68	.70	—	.77
Accepts “no”	.71	.68	.70	—	.78
Respects others	.79	.78	.80	—	.88
Shares	.76	.75	.77	—	.84
Apologizes	.79	.79	.80	—	.83
Is kind	.77	.76	.81	—	.85
School Readiness					
Understands words	.70	.72	—	.75	.63
Converses	.76	.77	—	.83	.65
Persists with tasks	.79	.82	—	.67	.77
Hygiene skills	.67	.67	—	.67	.65
Understands sentences	.69	.73	—	.71	.63
Listens to others	.75	.76	—	.75	.71
Pays attention	.86	.89	—	.72	.83
Listens to stories	.82	.85	—	.68	.80
Follows directions	.83	.87	—	.72	.80
Retells stories	.76	.79	—	.79	.68
Uses details	.75	.77	—	.81	.66
Works independently	.69	.70	—	.62	.68
Uses numbers/colors	.64	.67	—	.64	.58
Social Confidence					
Is self-confident	.70	.69	.65	.78	—
Acknowledges feelings	.66	.66	.65	.72	—
Asks for help	.65	.65	.62	.71	—
Stands up for self	.51	.51	.42	.62	—
Accepts closeness	.68	.66	.72	.69	—
Identifies feelings	.80	.81	.77	.84	—
Makes friends	.83	.84	.84	.84	—
Asks other to play	.79	.79	.77	.83	—
Enthusiastic about life	.73	.71	.74	.77	—
Family Involvement					
Sense of belonging	.63	—	.65	.64	.60
Trusts Sign. person	.62	—	.63	.63	.58
Positive relationships	.72	—	.74	.73	.69
Positive with parents	.69	—	.72	.68	.68
Involved in discussions	.75	—	.75	.77	.74
Positive with siblings	.70	—	.75	.64	.70
Participates in activities	.74	—	.75	.75	.73

Table 4. Omega, Omega Hierarchical, and Explained Common Variance Estimates.

	ω	ω_h	ECV (%)
Emotional Regulation	.97	.23	8
School Readiness	.96	.13	5
Social Confidence	.93	.13	3
Family Involvement	.95	.29	6
Strength Index	.99	.93	78

ω = omega; ω_h = omega hierarchical; ECV = explained common variance

variance, suggest high reliability given values ranging from .92 to .99; ω, like alpha, ranges from 0 (*completely unreliable*) to 1 (*perfectly reliable*). Omega hierarchical estimates, which indicate the proportion of scale score variance that can be attributed to true score variance of a single factor *when controlling for the other factors*, were large for the strength index (.93) and relatively small for the subscales (range = .13–.29). The ω_h values suggest that item response variance accounted for by the general factor is highly reliable, but that the item response variance accounted for by the group factors is far less reliable when the effect of the general factor is removed. The low ω_h values for the subscales indicate that these scores are far less precise when considered as unique constructs (Reise, 2012).

Discussion

Overall, the hypothesized internal structure of the assessment was confirmed, for children in the Head Start population, by the four-factor CFA model. In comparison with the four-factor model, the one-factor model demonstrated poor fit to the data indicating that the variances of item responses cannot be explained by a single, general latent factor. As the fit of the four-factor model demonstrated a statistically significant improvement over the one-factor model, we concluded that the four group factors were more consistent with the observed item response variances. However, the four latent factors were highly correlated suggesting a common source of variance, so the bifactor model was fit with both a general factor and four group factors. As expected, the bifactor model demonstrated a statistically significant improvement in fit over the four-factor model indicating that this latent structure was also consistent with observed item responses.[1]

More specifically, the bifactor model indicated a latent structure with a strong general factor and four *weaker* group factors. In other words, the bifactor model provided strong empirical support for the use of the overall strength score, but also provided results that indicated a need for researchers and Head Start personnel to be cautious of trying to interpret the subscale scores as independent sources of information about the behavioral and emotional strengths of young children in Head Start programs. Reliability estimates from the bifactor model also suggest that the PreBERS scores, for the total score and the subscales, are highly reliable measures of behavioral and emotional strengths; however, as noted above, ω_h estimates suggest that the majority of reliable variance in subscale scores is attributable to the general factor. The lack of reliable variance attributable to the group factors raises questions about using the subscale scores as reliable measures of the subdomain constructs.

Overall, the findings of the present investigation support the psychometric characteristics of the PreBERS scores, and we suggest that, for the time being, the assessment be used as the developers intended. However, there are a number of questions which still remain of which the most important question is "would it be best for researchers or practitioners to compute and interpret subscales scores or a single overall score?" While there are no established guidelines for drawing a conclusion about the univocal (i.e., unidimensional) nature of assessment scores, Reise, Moore and Haviland. (2010) and Reise (2012) suggest that when ω_h estimates (or ECV estimates) for subscale scores are as low as observed in this study, that assessment scores should likely be treated as univocal (i.e., as representing a single construct rather than distinct subdomains). On the other hand, the lack of strict item-parameter invariance does not necessarily support the univocal structure of the PreBERS ratings. Although there seems to be a slight preponderance of evidence supporting the use of overall PreBERS score rather than subscale scores, the question about how best to score and use the assessment data cannot be completely addressed by PreBERS data alone; instead, the decision to treat the scores as unidimensional needs to be carefully considered within the context of how factor scores relate to criterion measures of behavior and how practitioners or researchers intend to use the information. Future research should address the external validity of the latent structures to help inform decisions regarding how best to score the PreBERS assessment.

Limitations

Several limitations need to be noted. First, while the national sample of Head Start children was adequate and representative with respect to ethnicity, race, and disability status, the sample was not representative of Head Start children nationwide (see O'Brien et al., 2002) with respect to geographical region and age. The sample was overrepresented from the south (39% sample vs. 25% actual). In addition, the sample was somewhat overrepresented with 5 years old (23% sample vs. 8% actual). When the test developers re-norm the PreBERS, they should attempt to secure a more representative sample of children in Head Start settings particularly with respect to geographical region and age. Second, the PreBERS was developed based on data provided by preschool teachers and did not include data from parents or other caregivers. Thus, this study provides information on the PreBERS ratings of Head Start teachers and does not inform us on the ratings or expectations of parents of children in Head Start settings. Researchers (e.g., Achenbach, McConaughy, & Howell, 1987) have documented significant rating differences across informants (i.e., parents and teachers); thus, it is important to collect data on parents of Head Start children to develop a separate set of norms and to determine if ratings differ across teachers and parents. Third, the national sample of preschoolers was not randomly selected. Individuals were contacted through a variety of means and asked to assist. For this reason, the sample was a convenience sample of program directors and teachers who agreed to participate and thus the nature of the sample may have influenced the results. As such, the data do not inform us about children whose directors and teachers did not agree to participate and thus may have led to rater bias. A final limitation is related to the lack of information on the clustering of children within teachers (i.e., multiple students being rated by the same teacher) or within Head Start programs. As neither piece of information was collected as part of the PreBERS normative process, the clustering could not be accounted for in the CFA models.

While the present study demonstrated the presence of the four factors and the general strength index among this population, further research needs to be conducted on the reliability and validity of the PreBERS scores with Head Start children. First, as stated earlier, investigators need to determine the cross informant reliability of the PreBERS ratings, specifically the agreement between Head Start educational staff and parents. This will determine whether teachers and parents who interact with the child across settings view the child's strengths in a similar or dissimilar manner. Third, the item-level CFA model was demonstrated for children in Head Start settings but did not examine the invariance of measurement properties between males and females, across different ethnic or racial backgrounds, or between children with or without disabilities. Fourth, convergent validity studies with children in Head Start programs need to be conducted with other instruments of childhood functioning. For example, researchers should examine the association between the PreBERS scores and scores from tests of family cohesion, language and literacy, social interactions, among others. In addition, investigators need to use concurrent and predictive validation approaches in determining the usefulness of bifactor models of behavioral and emotional strengths. Finally, investigators need to conduct longitudinal studies of the PreBERS to evaluate the sensitivity to change of the scores.

Implications

Demonstrating evidence of reliability and validity of the PreBERS scores among a large sample of children enrolled in Head Start settings is important because this assessment can be used by Head Start personnel to identify children who may benefit from extra educational and therapeutic intervention to maximize their behavioral strengths, and set children up for academic success in Kindergarten and beyond. With 15% to 30% of Head Start children likely to be identified with emotional and behavioral problems (Qi & Kaiser, 2003), educators have a pressing need for a quick and reliable assessment that can identify individual child strengths using behavioral subscales (i.e., item content) that complement those addressed in Head Start classes. One of the primary purposes of Head Start is to maximize Academic Readiness, which is one of the subscales of the PreBERS. The other PreBERS subscales complement this primary Head Start focus by identifying individual child skills (Emotional Regulation, Social Confidence) and community strengths (Family Involvement) to support long-term academic success. Some of the strongest predictors of behavior problems by late elementary years include parenting stress and family psychopathology and parents' reports of child internalizing problems in preschool (Ashford, Smit, van Lier, Cuijpers, & Koot, 2008), which correspond to PreBERS subscales of positive strengths in Family Involvement and Emotional Regulation. If these early strengths could be effectively identified and fostered through preventative interventions, up to 57% of behavioral impairments at age 11 years could be ameliorated (Ashford et al., 2008).

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

Funding — The authors received no financial support for the research, authorship, and/or publication of this article.

Note

1. The bifactor model has 36 fewer degrees of freedom than the four-factor model (i.e., the bifactor model is more complex); therefore, we would expect the model fit to improve over that of the four-factor model as a function of the additional complexity and does not necessarily represent a substantive improvement in fit to the data.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213-232.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms & profiles: An Integrated System of Multi-Informant Assessment; Child Behavior Checklist for Ages 1½–5*. Burlington: University of Vermont, Research Center for Children, Youth, and Families.

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

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

Ashford, J., Smit, F., van Lier, P. A. C., Cuijpers, P., & Koot, H. M. (2008). Early risk indicators of internalizing problems in late childhood: A 9-year longitudinal study. *Journal of Child Psychology and Psychiatry*, *49*, 774-780.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling*, *7*, 461-483.

Briggs-Gowan, M., & Carter, A. S. (1998). Preliminary acceptability and psychometrics of the Infant–Toddler Social and Emotional Assessment (ITSEA): A new adult-report questionnaire. *Infant Mental Health Journal*, *19*, 422-445.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory–II: G is the key. *Psychological Assessment*, *25*, 136-145.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

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

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*, 796-846.

Carter, A. S., Briggs-Gowan, M. J., & Davis, N. O. (2004). Assessment of young children's social-emotional development and psychopathology: Recent advances and recommendations for practice. *Journal of Child Psychology and Psychiatry*, *45*, 109-134. doi:10.1046/j.0021-9630.2003.00316.x

Cress, C. J., Epstein, M. H., & Synhorst, L. (2010). Preschool Behavioral and Emotional Rating Scale (PreBERS): Test-retest and inter-rater reliability for children with disabilities. *Exceptionality*, *18*, 28-69.

Epstein, M. H. (2004). *Behavioral and Emotional Rating Scale—Second edition: A strength-based approach to assessment*. Austin, TX: Pro-Ed.

Epstein, M. H., & Synhorst, L. (2009). *Preschool Behavioral and Emotional Rating Scale*. Austin, TX: Pro-Ed.

Epstein, M. H., Synhorst, L. L., Cress, C. J., & Allen, E. A. (2009). Development and standardization of a test to measure the emotional and behavioral strengths of preschool children. *Journal of Emotional and Behavioral Disorders*, *17*, 29-37.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581-586.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling*, *7*, 251-270.

Greenspan, S. I., & Meisels, S. J. (1996). Toward a new vision of developmental assessment of infants and young children. In S. J. Meisels & E. Fenichel (eds.), *New visions for the developmental assessment of infants and young children* (pp. 11-26). Washington, DC: Zero to Three.

Griffith, A. K., Duppong Hurley, K., Trout, A. L., Synhorst, L., Epstein, M. H., & Allen, E. (2010). Assessing the strengths of young children at risk: Examining use of the Preschool Behavioral and Emotional Rating Scale with a head start population. *Journal of Early Intervention*, *32*, 274-285.

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.

Kauffman, J. M., & Landrum, T. J. (2009). *Characteristics of emotional and behavioral disorders of children and youth* (9th ed.). Upper Saddle River, NJ: Prentice Hall.

LeBuffe, P. A., & Naglieri, J. A. (1998). *The Devereux Early Childhood Assessment*. Lewisville, NC: Kaplan.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McCardle (eds.), *Contemporary psychometrics: A festschrift to Roderick P. McDonald* (pp. 275-340). Hillsdale, NJ: Lawrence Erlbaum.

McCain, M. N., Mustard, J. F., & Shanker, S. (2007). *Early Years Study 2: Putting science into action*. Toronto, Ontario, Canada: Council for Early Child Development.

McDonald, R. P. (1978). Generalizability in factorable domains: "Domain validity and generalizability." *Educational and Psychological Measurement*, *38*, 75-79.

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

Meisels, S. J., & Atkins-Burnett, S. (2000). The elements of early childhood assessment. In J. P. Shonkoff & S. J. Meisels (eds.), *Handbook of early childhood intervention* (2nd ed., pp. 231-257). New York, NY: Cambridge University Press.

Moss, J. (2006). *Child preference indicators: A guide for planning*. Oklahoma City: University of Oklahoma Health Sciences Center, Center for Learning and Leadership.

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

Nordness, P. D., Epstein, M. H., & Synhorst, L. (2009). Convergent validity and test-retest reliability of the Preschool Behavioral and Emotional Behavior Rating Scale: Parents as respondents. *Early Childhood Services*, *3*, 51-60.

O'Brien, R. W., D'Elio, M., Vaden-Kiernan, M., Magee, C., Younoszai, T., & Keane, M. J. (2002). *A descriptive study of head start families: FACES Technical Report I*. http://www.acf.hhs.gov/programs/opre/hs/faces/reports/technical_report/technical_report_pdf

Powell, D., Dunlap, G., & Fox, L. (2006). Prevention and intervention for the challenging behaviors of toddlers and preschoolers. *Infants & Young Children*, *19*, 25-35.

Qi, C. H., & Kaiser, A. P. (2003). Behavior problems of preschool children from low-income families: Review of the literature. *Topics in Early Childhood Special Education*, *23*, 188-216.

Raykov, T. (1997). Simultaneous study of individual and group patterns of latent longitudinal change using structural equation modeling. *Structural Equation Modeling*, *4*, 212-236.

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

Reise, S. P., Horan, W. P., & Blanchard, J. L. (2010). The challenges of fitting an item response theory model to the Social Anhedonia Scale. *Journal of Personality Assessment*, *93*, 213-224.

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.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2013). *Assessment in special and inclusive education* (12th ed.). Belmont, CA: Wadsworth.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120.

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

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1998). *Vineland Social-Emotional Early Childhood Scale*. Circle Pines, MN: American Guidance Service.

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

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

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.