School Psychology Quarterly
2017, Vol. 32, No. 2, 226–239

© 2016 American Psychological Association
1045-3830/17/$12.00 http://dx.doi.org/10.1037/spq0000155

Structural Validity of CLASS K–3 in Primary Grades: Testing Alternative Models

Lia E. Sandilos
University of Virginia

Sarah Wollersheim Shervey
University of South Dakota

James C. DiPerna, Puiwa Lei, and Weiyi Cheng
The Pennsylvania State University

This study examined the internal structure of the Classroom Assessment Scoring System (CLASS; K–3 version). The original CLASS K–3 model (Pianta, La Paro, & Hamre, 2008) and 5 alternative models were tested using confirmatory factor analysis with a sample of first- and second-grade classrooms (N = 141). Findings indicated that a slightly modified version of the original CLASS K–3 3-factor model best fit the current data. Although stable findings emerged across the current and previous studies, particularly in relation to the presence of 3 latent domains, there is also some variability across structures at different grade levels with regard to the bifactor and 3-factor models.

Keywords: CLASS K–3, structural validity, teacher–child interactions

An ecological perspective of the classroom environment posits that dynamic interactions in the education setting influence children's schooling experience and ultimately their performance (Rimm-Kaufman & Pianta, 2000). In particular, positive dyadic exchanges between children and adults have the potential to foster children's development (Hamre, Hatfield, Pianta, & Jamil, 2014). Prior research has indicated that interactions between teachers and students in early childhood and elementary classrooms may be critical for students' academic and social-emotional outcomes (e.g., McCormick & O'Connor, 2015; Rudasill, Reio, Stipanovic, & Taylor, 2010). Thus, the development of measures that capture the nature of these teacher–child interactions has been a focal area in educational research for many years. Empirical work on this topic has identified systematic classroom observations as a practical and potentially effective way to assess these interactions (Zaslow, Martinez-Beck, Tout, & Halle, 2011). The Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) is one such observational instrument that has been widely used in both research and practice to measure teacher-student interaction quality.

The CLASS is grounded in theory and reflects research regarding high quality practices in classroom settings. The primary domains assessed within the prekindergarten and early elementary (K–3) versions[1] of CLASS are Emotional Support, Classroom Organization, and Instructional Support. The Emotional Support domain examines the teacher's ability to foster a warm and positive climate in which the students can exercise autonomy and the teacher is

This article was published Online First May 12, 2016.

Lia E. Sandilos, Curry School of Education, University of Virginia; Sarah Wollersheim Shervey, Department of Counseling and Psychology in Education, University of South Dakota; James C. DiPerna, Puiwa Lei, and Weiyi Cheng, Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090438 to The Pennsylvania State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Lia E. Sandilos, Curry School of Education, University of Virginia, 268 Ruffner Hall, P.O. Box 800784, Charlottesville, VA 22904. E-mail: les2h@virginia.edu

[1] Separate versions of CLASS are available for pre-k through secondary classrooms.

sensitive to students' academic and emotional needs. The Classroom Organization domain assesses the teacher's skill at managing student behaviors, establishing routines, and using varied modalities for learning. The Instructional Support domain examines the teacher's ability to provide constructive feedback and scaffolding, model novel vocabulary, extend student responses, and foster analytical thinking skills (Pianta et al., 2008).

In a research context, CLASS has been widely used when exploring domain-general (global) teaching quality and teacher–child interactions (e.g., LoCasale-Crouch et al., 2007). The CLASS has also been adopted into prekindergarten educational practices. Head Start uses CLASS scores to help determine the accreditation of new prekindergarten centers around the nation (Hamre et al., 2014). In addition, teacher professional development programs (i.e., MyTeachingPartner) and other preschool curricular materials (e.g., MyTeachingPartner Math/Science) have been developed based on the CLASS framework. Despite its rapid emergence as a popular measure of teacher–child interactions, the links between CLASS scores and child outcomes have yielded primarily low to moderate relations (e.g., Burchinal et al., 2008; Mashburn et al., 2008). These modest relations to outcomes, as well as international interest in the use of the measure with diverse classroom populations, have prompted the authors of the CLASS, as well as independent researchers both nationally and internationally, to further examine the internal structure of the measure (Hamre et al., 2014).

The original CLASS framework was based on data collected in over 4,000 prekindergarten, kindergarten, first-grade, third-grade, and fifth-grade classrooms across the United States (Hamre, Pianta, Mashburn, & Downer, 2007). Using comparative factor analysis (CFA), a three-factor model (Emotional Support, Classroom Organization, and Instructional Support) demonstrated the best overall fit in prekindergarten through third-grade classrooms (Hamre et al., 2007). Further, the CLASS framework has strong theoretical and conceptual underpinnings to support the three proposed factors (Hamre et al., 2007). The primary limitation to this initial validity study was that the fit indices did not consistently meet criteria for close fit across grades, suggesting potential error in the model (Browne & Cudeck, 1993; Hu & Bentler, 1999). As a result of gradual changes over time in the types and number of dimensions included within those three domains, direct comparisons of models across grade levels have not been feasible. The final published versions of the CLASS Pre-K and K–3 measures have three domains and 10 dimensions.

Pakarinen et al. (2010) published an international study of CLASS structural validity with a small sample of 49 Finnish kindergarten classrooms (Pakarinen et al., 2010). CFA findings revealed that the model demonstrating the best fit omitted the Negative Climate dimension from the Emotional Support domain altogether due to its low discriminant validity. Although the three-domain, nine-dimension model had the best fit with this sample, the resulting domains in the final model (Emotional Support, Classroom Organization, and Instructional Support) exhibited multicollinearity (>.90). The removal of Negative Climate and the high domain intercorrelations are potentially noteworthy as these findings may represent cultural differences in teacher–child interactions between U.S. and Finnish classrooms. Pakarinen et al. (2010) hypothesized that the differences found in the model may have been a product of the largely constructivist teaching approach used in Finnish classrooms, which integrates relationships and instruction even more closely than in U.S. classrooms. However, given the fact that small samples can result in unstable parameter estimates, it is difficult to know if the observed changes were a product of cultural differences or unique to the sample. Thus, replication of this particular model with a larger domestic sample is warranted.

Since its initial release, the authors of CLASS also have explored alternative models that differ from the original structure of the scale. For example, Hamre and colleagues (2014) tested a bifactor structure of CLASS with 325 prekindergarten classrooms. A bifactor model is useful to consider when indicators and/or factors are highly correlated. Within a bifactor structure, all of the indicators load onto one general factor. Select indicators also load onto orthogonal (uncorrelated) domain-specific factors that account for unique variance beyond the general factor (Chen, West, & Sousa, 2006). The bifactor model identified by Hamre and colleagues consists of a global factor, Responsive Teaching, as well as two domain factors, Positive Management and Routines (i.e., combined several Emotional Support and Classroom Organization dimensions) and Cognitive Facilitation (i.e., Instructional Support dimensions). With a prekindergarten sample, the bifactor model demonstrated improved fit over the original three-domain CLASS structure. As the authors indicated, the bifactor structure also presented an effective way to correct for multicollinearity among the domain scores in other studies because the factors are orthogonal.

The bifactor model also was tested by Madill (2014) with a sample of first-, third-, and fifth-grade classrooms. Though the Hamre et al. (2014) bifactor model did not fit the data well, an alternative bifactor model exhibited better fit. This bifactor model, based on earlier work by Jones, Molano, Brown, and Aber (2013), dropped Positive Climate from the Positive Management and Routines domain (subsequently renamed Management and Routines) and added the Instructional Learning Formats dimension to the Cognitive Facilitation domain (Madill, 2014).

Sandilos, DiPerna, and The Family Life Project Key Investigators (2014) tested the original three-domain CLASS model and the Hamre et al. (2014) bifactor structure with a sample of 417 kindergarten classrooms from rural, low-income classrooms in the northeastern and southeastern United States. Using CFA, the best-fitting model in this study moved the Behavior Management dimension from the Classroom Organization domain to the Emotional Support domain. The authors hypothesized that the strong relation between Behavior Management and Emotional Support may be linked to the developmentally appropriate emphasis on positive and emotionally supportive strategies for modifying behavior that is often present within early elementary classrooms.

Rationale

The Standards for Educational and Psychological Testing (American Education Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) emphasize the importance of internal structure when considering the validity of scores from an assessment. Although structural validity evidence is not sufficient to establish validity of the scores, it is essential because the indicators and constructs established through examinations of internal structure are used to generate scores from a test and to examine other forms of validity evidence such as test-criterion relationships, validity generalization, and evidence based on consequences of testing (AERA et al., 2014). Thus, it is critical to establish evidence of valid internal structure for any widely used measure.

The structural validity of CLASS has been tested across prekindergarten and elementary samples, and each study has identified some variations in the best-fitting structural model from those that preceded it. Moreover, some alternative models have been tested repeatedly and consistently exhibited poor fit (e.g., one-factor, two-factor), whereas other more promising models would benefit from replication (e.g., Hamre et al., 2014; Sandilos et al., 2014). In addition, researchers have used different fit indices, as well as criteria that vary in stringency, when determining best-fitting models. Thus, testing the aforementioned models (original and modified) on the same sample with consistent fit criteria would allow for direct comparison of the various structures emerging from previous studies of the CLASS.

Currently, there are few replication studies (psychometric or otherwise) in education research (e.g., Makel & Plucker, 2014). Thus, given the variations in CLASS structural models across previous studies, the primary aim of this study was to replicate the original and alternative CLASS structures with a sample of primary classrooms (Grades 1–2). The models (see Table 1) included both the original CLASS model (Pianta et al., 2008) and the alternative structures identified by Pakarinen et al. (2010); Hamre et al. (2014); Madill (2014), and Sandilos et al. (2014).

Method

Participants

Data for the current study were collected from 141 first- and second-grade classrooms across seven elementary schools in the mid-Atlantic region of the Unites States. Five of the elementary schools were from an urban district, and two elementary schools were from a small

Table 1
Classroom Assessment Scoring System (CLASS) Structural Validity Studies

Study	Sample	Data collection	Key findings	Fit indices

Pianta, La Paro, & Hamre (2008); Hamre, Pianta, Mashburn, & Downer (2007)	4,000 prekindergarten through 3rd grade classrooms across the U.S.	Across studies, data were collected with varying numbers of cycles and at different points in the year	Identified 3-factor, 10-dimension model of CLASS	SRMR = not reported
				RMSEA = .14
				CFI = .91

Pakarinen et al. (2010)	49 kindergarten classrooms in Finland	2 days of observations were aggregated, cycles ranged from 2–5 per day (1–2.5 hr); observations conducted February through April	Negative Climate removed	SRMR = .04
				RMSEA = .14
				CFI = .96

Hamre, Hatfield, Pianta, & Jamil (2014)	325 preschool and Head Start classrooms in 10 sites across the U.S.	Minimum of four 15-min cycles, (2.5–4 hr); observations conducted January through mid-March	Bifactor structure (Responsive Teaching, Positive Management & Routines, Cognitive Facilitation)	SRMR = .04
				RMSEA = .11
				CFI = .96

Madill (2014)	147 1st, 3rd, & 5th grade classrooms in rural, mid-size, and urban areas in northeastern and midwestern U.S.	1 observation per classroom; 4 cycles (2 hr); observations conducted within two months of 1st day of school	Modified bifactor structure (Positive Climate removed from Management & Routines; Instructional Learning Formats added to Cognitive Facilitation)	SRMR = not reported
				RMSEA = .11
				CFI = .97

Sandilos, DiPerna, & the Family Life Project Investigators (2014)	417 kindergarten classrooms in rural, low-income areas in northeastern and southeastern U.S.	1 observation per classroom, 2 cycles (1 hr); observations conducted October through December	Behavior Management moved to the Emotional Support domain	SRMR = .06
				RMSEA = .097
				CFI = .95

Note. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root-mean-square residual.

rural school district. Across the schools, 69.9% of students received free or reduced price lunch, and the racial/ethnic composition of the student population was approximately 65.8% Caucasian, 18.1% African American, 8.6% Hispanic, and 7.5% Other (i.e., Asian, Pacific Islander, Native American; National Center for Education Statistic, 2014). Approximately 20–25 students were enrolled in each participating classroom. The majority of participating teachers were female (89.4%), Caucasian (97.2%), and experienced in teaching (M = 14.4 years, SD = 9).

Measures

CLASS K–3. The CLASS K–3 assesses the overall quality of the classroom instructional environment in early elementary school. It is a structured observation system where trained observers rate teachers on 10 dimensions on a 7-point scale ranging from 1–2 (*low*), 3–5 (*middle*), to 6–7 (*high*). In the original CLASS model, the Emotional Support domain is made up of Positive Climate, Negative Climate, Teacher Sensitivity, and Regard for Student Perspectives dimensions. The Classroom Organization domain consists of Behavior Management, Productivity, and Instructional Learning Formats. The Instructional Support domain is comprised of Concept Development, Quality of Feedback, and Language Modeling. According to the authors, a minimum of two observation cycles should be completed, and observation cycles consist of 20 min of observation and note-taking followed by 10 min of scoring (Pianta et al., 2008).

CLASS scores remain relatively stable across four continuous cycles (Curby, Grimm, & Pianta, 2010), and correlations between two and four cycles in preschool and third grade are typically high (*r* = .87–.95; Pianta, La Paro, & Hamre, 2008). Pianta et al. (2008) reported moderate to high internal consistency coefficients (.63–.88) across two cycles for dimensions and domains in preschool and third grade. Internal consistency for the present sample was high as well, ranging from .81 to .93 for dimensions and domains. Acceptable interrater agreement within 1 point (i.e., adjacent agreement) must be 80% or higher (Pianta et al., 2008). In the current study, interrater agreement within 1 point was high, ranging from 91% to 99% across CLASS domains.

Procedure

Data were collected in the fall of the academic year by observers who achieved the mastery criteria (accuracy ≥80%) and were trained by CLASS-certified instructors. CLASS training and certification involves two rigorous days of studying the CLASS framework and coding videotaped observations. To pass the certification test, observers must achieve 80% reliability across five cycles and 80% reliability overall on each dimension. The certified CLASS observers consisted of two male and 11 female graduate students with extensive training in classroom observation and assessment. Throughout data collection, interrater agreement checks were conducted for 30% of the classroom observations.

The CLASS observation consisted of two 30-min cycles. CLASS dimensions were calculated by averaging scores across cycles within an observation. The dimension scores were used for the modeling analyses. Demographic data were collected using survey questions that teachers completed electronically.

Data Analysis

Mplus (Muthén & Muthén, 2008–2012) was used to conduct the confirmatory factor analyses of CLASS K–3 data. Cluster identifiers were specified to account for the data structure of classrooms being nested within schools. The default maximum likelihood estimator was used for model estimation. Model fit was evaluated by multiple fit indices (Hu & Bentler, 1999). To evaluate the overall fit of the model, root mean squared error of approximation (RMSEA) and the standardized root-mean-square residual (SRMR) were considered. The RMSEA is an absolute fit index that represents the lack of fit of the model to the population covariance matrix. RMSEA values less than .05 are considered indicative of a good fit, values between .05 and .08 as an adequate fit, values between .08 and .10 as a mediocre fit, and values greater than .10 are not acceptable (Browne & Cudeck, 1993). The standardized root-mean-square residual (SRMR) is an index based on covariance residuals. The SRMR is considered favorable if the value is less than or equal to .08 (Hu & Bentler, 1999), but values as high as .10 can be interpreted as acceptable (Schermelleh-Engel, Moosbrugger, & Muller, 2003).

The model comparison fit indices examined in this study were the comparative fit index (CFI), the Tucker-Lewis index (TLI), Akaike information criterion (AIC), and the Bayesian information criterion (BIC). The CFI and TLI are incremental fit indices that assess the improvement in fit of a proposed model relative to a baseline (null model). CFI and TLI values greater than or equal to .90 are regarded as evidence for an acceptable-fitting model (Bentler & Bonett, 1980; Hu & Bentler, 1999; Kline, 2013), whereas values greater than .95 are considered a good fit (Schermelleh-Engel et al., 2003). The AIC and BIC are predictive fit indices that assess model fit in terms of hypothetical replications from the same population. Lower values of AIC and BIC are preferred, indicating models are more likely to replicate (Kline, 2013). After testing the fit of each of the previously proposed models, modification indices were used to construct a model that best fit the data.

Five a priori models were tested in the current study. These included the original CLASS model (Hamre et al., 2007; Pianta et al., 2008) and the alternative structures identified by Pakarinen et al. (2010); Hamre et al. (2014); Madill (2014), and Sandilos et al. (2014). Because none of the models met all fit criteria, a sixth model was tested based on modification indices (Table 1; Figures 1–2).

Results

Descriptive statistics are reported in Table 2. Correlations among dimensions ranged widely from .01–.73. Dimension means ranged from

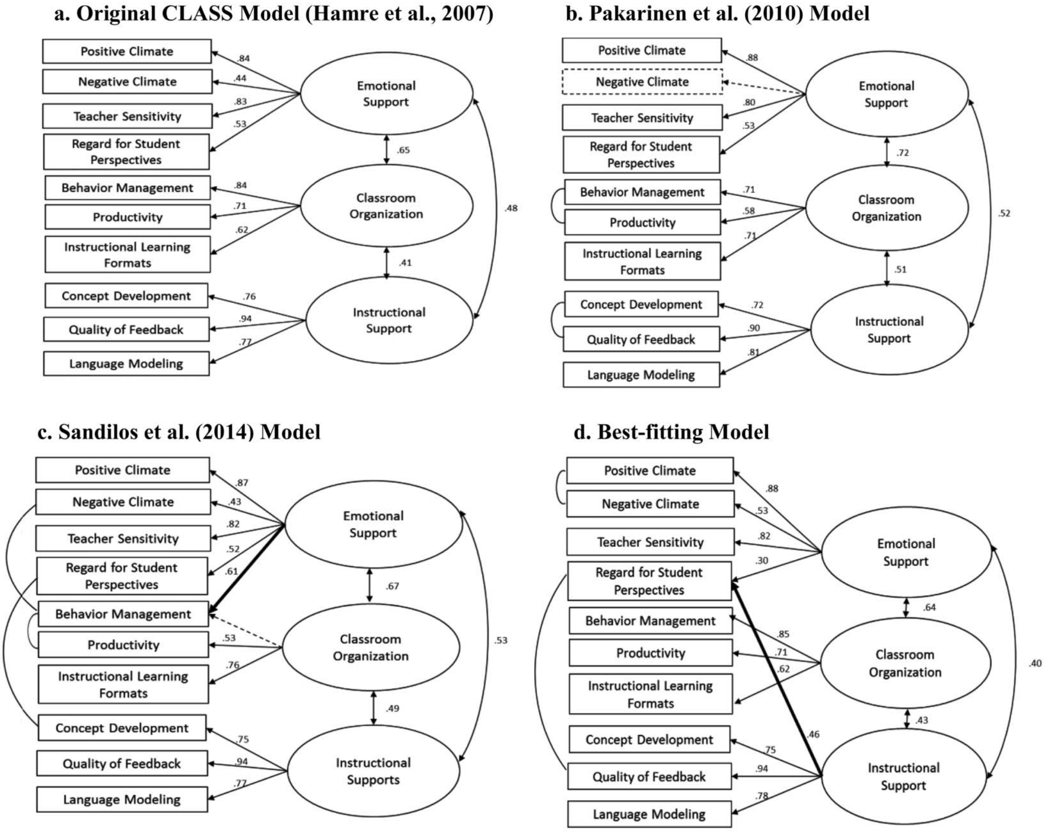

Figure 1. Loadings for Classroom Assessment Scoring System (CLASS) structural models with three latent domains. Dotted lines represent the removal of a pathway or dimension relative to the original CLASS model. Bold lines represent the addition of a pathway relative to the original CLASS model. All reported loadings are significant ($p < .05$).

2.51 to 6.66. On average, Emotional Support and Classroom Organization dimension scores fell within the middle to high ranges, while Instructional Support dimension scores fell within the lower range. Many previous studies using prekindergarten and early elementary samples have also found teachers' scores on the Instructional Support domain to be skewed to the lower end of the scale, as compared to the other two domains (e.g., Curby, Rimm-Kaufman, & Ponitz, 2009; Hamre et al., 2014).

Three-Factor Models

The original CLASS K–3 structure of 10 dimensions and three domains (Hamre et al., 2007; Pianta et al., 2008; Figure 1a) was tested first, and multiple fit indices did not meet a priori thresholds for adequate fit (see Table 3). Specifically, RMSEA was inflated above the .10 threshold and TLI fell below the minimum criterion of .90.

Next, the model from the Pakarinen et al. (2010) study (Figure 1b) was tested. In this model, Negative Climate is removed from the original CLASS structure, and two pairs of residual errors are correlated (Behavior Management & Productivity, Quality of Feedback & Concept Development). This model also did not fit the data well (see Table 3), however, as none of the fit indices met even the least-restrictive fit criteria.

The third model tested was the structure identified by Sandilos et al. (2014). As shown in Figure 1c, the primary changes in this model relative to the original CLASS structure consist

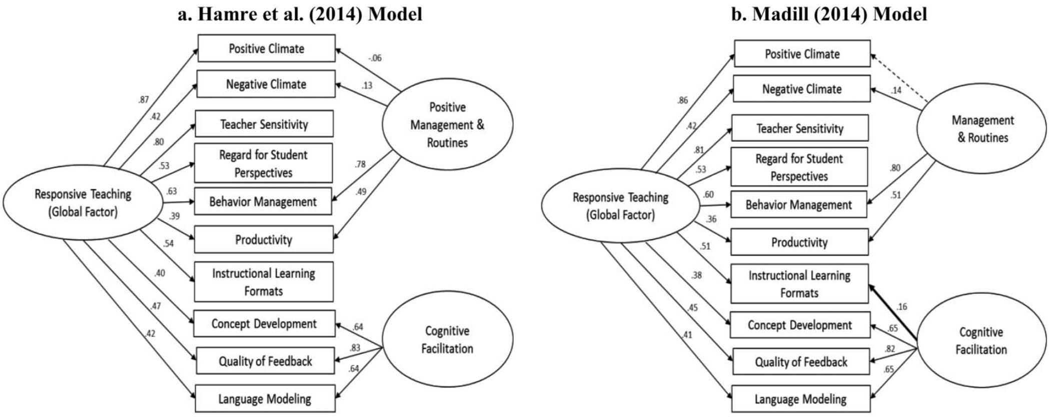

Figure 2. Factor loadings for Classroom Assessment Scoring System (CLASS) bifactor models. Dotted lines represent the removal of a pathway or dimension relative to the Hamre et al. bifactor model. Bold lines represent the addition of a pathway relative to the Hamre et al. (2014) bifactor model. All reported loadings are statistically significant ($p < .05$), with the exception of Positive Climate in the Hamre et al. (2014) model.

of adding a direct pathway from Emotional Support to Behavior Management, and removing the pathway from Classroom Organization to Behavior Management. In addition, the residuals of Productivity and Behavior Management, Behavior Management and Negative Climate, and Regard for Student Perspectives and Concept Development are correlated. Test of the Sandilos et al. model with the current sample yielded indices falling within the mediocre to good fit range (see Table 3 and 4).

Bifactor Models

Two bifactor structures also were tested as part of this study. The Hamre et al. (2014) bifactor structure (Figure 2A) features three uncorrelated factors: Responsive Teaching (global factor), Positive Management and Routines (domain-specific factor), and Cognitive Facilitation (domain-specific factor). When testing the Hamre et al. bifactor structure, the Positive Climate dimension did not significantly load on the Positive Management and Routines factor. As such, the revised bifactor model specified by Madill (2014) also was tested (Figure 2b). This model, which omits Positive Climate from the Positive Management and Routines factor and includes Instructional Learning Formats on the Cognitive Facilitation factor, improved fit relative to the Hamre et al. model. However, the

Table 2
Descriptive Statistics for Classroom Assessment Scoring System Dimensions

Dimension	1	2	3	4	5	6	7	8	9	10
1. Positive Climate										
2. Negative Climate	.32***									
3. Teacher Sensitivity	.72***	.43***								
4. Regard for Student Perspectives	.44***	.29***	.40***							
5. Behavior Management	.51***	.37***	.50***	.25**						
6. Productivity	.27**	.16	.25*	.25**	.62***					
7. Instructional Learning Formats	.44***	.18*	.33***	.42***	.47***	.49***				
8. Concept Development	.34***	.01	.23**	.41***	.25**	.26**	.36***			
9. Quality of Feedback	.39***	.07	.34***	.42***	.28***	.26**	.36***	.71***		
10. Language Modeling	.38***	.04	.26*	.48***	.21*	.24**	.33***	.57***	.73***	
M	5.18	6.66	4.77	3.72	5.65	5.42	4.43	2.56	2.92	2.51
SD	1.18	.75	1.20	1.13	1.17	1.06	1.21	1.15	1.24	1.18

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3
Fit Statistics for Classroom Assessment Scoring System Structural Models

Models	χ^{2a}	*df*	CFI	TLI	RMSEA	SRMR	AIC	BIC
Pianta et al. (2008); Hamre et al. (2007) (Original Model)	96.67	32	.909	.872	.120	.081	3,805.013	3,902.322
Pakarinen et al. (2010)[b]	115.72	31	.881	.827	.139	.112	3,825.514	3,925.772
Hamre et al. (2014)	98.55	29	.902	.848	.130	.083	3,813.514	3,919.670
Madill (2014)	93.09	29	.910	.860	.125	.081	3,809.965	3,916.121
Sandilos et al. (2014)	67.89	29	.945	.915	.098	.071	3,792.509	3,898.664
Best-fitting model	59.13	29	.958	.934	.086	.063	3,783.109	3,889.264

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root-mean-square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion.
[a] All chi-square tests are significant ($p < .05$). [b] Omitted Negative Climate dimension.

RMSEA for the Madill et al. model still exceeded the .10 threshold and the TLI fell below the minimum criterion of .90 (see Table 3). For both bifactor models, Behavior Management exhibited a nonsignificant (negative) residual variance that needed to be fixed to zero in order for the models to converge. The presence of this negative residual variance may indicate error in the bifactor structure with the current sample (Muthén & Muthén, 2008-2012).

Best-Fitting Model

Finally, because none of the tested CFA models yielded fit indices meeting all of the good fit thresholds, an alternative best-fitting model was developed using modification indices. A CFA model building and trimming approach (Kline, 2013) was applied to the original CLASS model to identify a best-fitting factor structure for the current sample.[2] Specifically, three changes were made to the original model (Figure 1d). First, a direct path was added from Instructional Support to Regard for Student Perspectives. Second, the residuals of Positive Climate and Negative Climate were correlated. Finally, the residuals of Quality of Feedback and Regard for Student Perspectives also were correlated (see Table 3). Though RMSEA was slightly above the .08 threshold for acceptable fit, this model fit the data best across indices and yielded the least substantive modifications to the original model.

Discussion

The purpose of this study was to examine the internal structure of CLASS K–3 (Pianta et al., 2008). Previous factor analytic studies of CLASS have examined three-factor (Hamre et al., 2007; Pakarinen et al., 2010; Sandilos et al., 2014) and bifactor (Hamre et al., 2014; Madill, 2014) models. Although these latter (bifactor) models have emerged recently as a possible alternative structure that accounts for strong relations among CLASS domains, neither the Hamre et al. (2014) model nor the Madill (2014) model demonstrated adequate fit in the current sample of first- and second-grade classrooms. It is important to note, however, that the Hamre et al. model originally was tested with a prekindergarten sample, and the Madill model was tested with a sample that included first-, third-, and fifth-grade classrooms. Thus, it is possible that a bifactor model is a more appropriate structure for the CLASS when used in prekindergarten or intermediate (Grades 3–5) classrooms. However, for the present sample of first- and second-grade classrooms, a three-factor structure fit the data best.

Specifically, both the three-factor model originally identified by Sandilos and colleagues (2014) with a kindergarten sample and the final three-factor model from the current study demonstrated improved fit over the bifactor structure based on the TLI and RMSEA indices. Because TLI is a measure of the likelihood of model replication, this index tends to improve with model parsimony (Kline, 2013), which is an advantage of the three-factor models relative to the bifactor models. The RMSEA is an indicator of overall error in the model (Kline, 2013),

[2] An alternative best-fitting bifactor model was also explored; however, none of the modification yielded a better fit than the best-fitting three-factor stucture.

Table 4
Loading of Dimensions on Emotional Support, Classroom Organization, and Instructional Support Across Studies Supporting a Three-Factor Classroom Assessment Scoring System Model

Dimension	Pianta et al. (2008)	Pakarinen et al. (2010)	Sandilos et al. (2014)	Current study
Positive Climate	ES	ES	ES	ES
Negative Climate	ES	—	ES	ES
Teacher Sensitivity	ES	ES	ES	ES
Regard for Student Perspectives	ES	ES	ES	**ES/IS**
Behavior Management	CO	CO	**ES**	CO
Productivity	CO	CO	CO	CO
Instructional Learning Formats	CO	CO	CO	CO
Concept Development	IS	IS	IS	IS
Quality of Feedback	IS	IS	IS	IS
Language Modeling	IS	IS	IS	IS

Note. ES = emotional support; CO = classroom organization; IS = instructional support. Bold indicates variation from other studies.

and the lower RMSEA values for the three-factor models indicate that this structure demonstrates a better fit with the current data than a bifactor structure. It is important to note that the RMSEA values found in both the Hamre et al., (2014) and Madill (2014) bifactor studies (.11–.13) were consistent with the values found in the current study (.125–.130); however, these values exceed recommended thresholds for the RMSEA index (≤.08; Browne & Cudeck, 1993; Kline, 2013).

The three-factor models tested in the current study consisted of the original model (Pianta, Hamre, & La Paro, 2008) and alternative models by Pakarinen et al. (2010) and Sandilos et al. (2014). The Pakarinen et al. model exhibited the poorest fit with the current data. This finding is not surprising given the model initially was identified with a very small classroom sample, which can increase error, result in unstable parameter estimates, and decrease the likelihood of replication. Alternatively, cultural differences may result in structural changes when the CLASS is used outside of the United States. Pakarinen et al. (2010) suggested that the instructional behaviors typically reflected within the Negative Climate dimension (e.g., punitive language, sarcasm, eye-rolling) may be less common in Finnish classrooms given their particularly strong emphasis on positive teacher–child relationships. In addition, Määttä and Uusiautti (2012) found that Finnish teachers are generally pleased with their jobs (e.g., feel respected for profession, content with salary, etc.); whereas U.S. teachers have reported significant levels of burnout and stress for several decades (Kyriacou, 2001; Whitaker et al., 2013). The behaviors captured by the Negative Climate dimension may not have occurred in Finnish classrooms, in part, because job stress and dissatisfaction are less prevalent among Finnish teachers than U.S. teachers. Given the small sample size of the Pakarinen study and poor fit of that model with the current sample, there is insufficient evidence to justify use of this model for U.S. classrooms. In addition, more studies need to be conducted with larger international classroom samples before the Pakarinen et al. (2010) model can be considered further for use in international research or practice.

Though the Sandilos et al. (2014) three-factor model did not result in the best fit with the present data, a greater number of the fit indices for this model exceeded the a priori criteria than those of the Pakarinen model or the original three-factor model. The substantive modification in this model (i.e., placement of Behavior Management on the Emotional Support domain) relative to the original CLASS structure reflects the strong relations between Emotional Support and Classroom Organization during the early years of schooling. Sandilos and colleagues originally identified this model with a sample of kindergarten classrooms. As they noted, kindergarten teachers often spend time teaching children strategies for modifying their own behavior by learning to understand, regulate, and clearly express emotions (Sandilos et al., 2014), which could account for the link between Behavior Management and Emotional Support in kindergarten but may not generalize to later grade levels.

The best-fitting model in the current study maintained a three-factor structure; however, Regard for Student Perspectives loaded significantly onto both its original domain (Emotional Support) as well as an additional domain (Instructional Support). This structural change suggests that Regard for Student Perspectives, which encompasses support for student expression and autonomy, may reflect both instructional and emotional aspects of teaching quality. When teachers engage in behaviors that result in higher scores on this dimension, students may feel supported interpersonally and a stronger connection to the instructional content. Curby and colleagues (2013) also found associations between Emotional Support and Instructional Support domains when examining CLASS K–3 data in third- and fourth-grade classrooms. They hypothesized that teachers may need to be emotionally sensitive to students' learning needs as academic content becomes more difficult. Thus, providing students with opportunities to express themselves and exhibit autonomy (i.e., Regard for Student Perspectives) allows teachers to get to know individual students better which, in turn, may lead to more individualized or differentiated instructional techniques on the part of the teacher (Curby, Rimm-Kaufman, & Abry, 2013).

Although there is variability across the models that have emerged from the six studies of the CLASS structure, there are a number of latent variables and loadings that are stable as well. Notably, Quality of Feedback, Concept Development, and Language Modeling dimensions have consistently loaded together onto one latent factor, which is referred to as Instructional Support in the three-factor structures and as Cognitive Facilitation in the bifactor structures of Hamre et al. (2014) and Madill (2014). Two dimensions loading onto Emotional Support, Positive Climate and Teacher Sensitivity, also have been stable across three-factor studies. In addition, Productivity and Instructional Learning Formats have consistently loaded together onto the Classroom Organization domain across three-factor models. These patterns indicate stability in the relations between these dimensions and their underlying domains across grades and classrooms. Overall, findings generally support a three-domain latent structure of CLASS in the primary grades (K–2).

The dimensions that form the Emotional Support and Classroom Organization domains in the original three-factor structure (Pianta et al., 2008) also have displayed some variability across studies. Specifically, Behavior Management has shifted from Classroom Organization to Emotional Support (Sandilos et al., 2014), Negative Climate has been removed from Emotional Support (Pakarinen et al., 2010), and Regard for Student Perspectives has dually loaded on Classroom Organization and Instructional Support (best fitting-model from present study). Moreover, the latent factors of the bifactor models deviate from the original three-factor structure in that some of the dimensions contributing to the Emotional Support and Classroom Organization domains in the original CLASS K–3 model are combined into one latent factor, (Positive) Management and Routines, and all dimensions load onto a general latent factor (Responsive Teaching) as well. The variability of these dimensions across domains could reflect developmental shifts in instructional practices across grade levels. Variations in structure also may be more substantive at the pre-k and intermediate (Grades 3–5) levels based on the support for bifactor model at these levels (Hamre et al., 2014; Madill, 2014, respectively). However, these particular variations in structure must be replicated across grade levels before it can be determined if these variations are due to sampling error or actual differences in relationships resulting from age-or-developmental differences in classroom instruction.

One additional potential explanation for variability in the CLASS factor structure across studies could stem from the method used to assign ratings to the dimensions. Each of the 10 CLASS dimensions encompasses multiple teacher behaviors. For example, when evaluating the Positive Climate dimension, observers are expected to consider a variety of teacher behaviors such as using a warm tone and affect when interacting with students, communicating with students using praise and respectful language, and maintaining close proximity with students. Despite the rigorous training requirements for CLASS, it is possible that individual observers may weigh some behaviors more heavily than others when assigning their overall rating for a dimension. Thus, the current format of CLASS may be more open to observer subjectivity because it does not require each observer to rate each characteristic individually. Though error is difficult to eliminate in any measure that requires human judgment, one way to potentially reduce variability across observers in future versions of CLASS would be to require that observers rate each teacher behavior within a dimension and then aggregate all of the individual ratings to create the score for that dimension.

Data from the current study support the presence of three latent domains (Emotional Support, Classroom Organization, and Instructional Support) with dimension loadings that are fairly consistent with the original CLASS model. However, when using CLASS and interpreting such observational data, it is important for researchers and practitioners to be cognizant that the three latent factors are related and may have some overlap. For example, the best-fitting model from the current study revealed links between Emotional Support and Instructional Support through the dual loading of Regard for Student Perspectives onto both domains. In a practical context, this modification suggests that it may be important for educators to incorporate students' opinions and background experiences in an effort to differentiate instruction and strengthen the emotional climate in the classroom. Moreover, developmentally appropriate instruction for first and second-grade students should incorporate more opportunities for autonomy and self-expression, which are key aspects of Regard for Student Perspectives. Eccles (1999) posited that middle childhood (ages 6–10) is a critical period for increased levels of student autonomy, competence, and self-expression. As a result, both Emotional Support and Instructional Support likely play a key role in demonstrating Regard for Student Perspectives at this developmental level; whereas these domains appear to remain more distinct in kindergarten and preschool.

In addition, two of the replicated models (Pakarinen et al., 2010; Sandilos et al., 2014) and the best-fitting model in this study allowed for correlated residuals. Correlated residuals occur when there is shared variance between dimensions that is not explained by the underlying latent construct. For the best-fitting model, correlated residuals were allowed between Positive and Negative Climate because these dimensions load onto the same domain and may reflect opposite ends of the classroom climate continuum. Thus, despite the fact that they are considered unique dimensions, observers may be scoring these dimensions with a single continuum in mind, which could result in additional shared variance beyond that explained by the latent factor (Emotional Support). The residuals of Quality of Feedback and Regard for Student Perspectives also were allowed to correlate because both dimensions appear to reflect instructional interactions that require a respect for students and student-centered instruction. For example, teachers with higher scores on Regard for Student Perspectives will encourage student talk, elicit student perspectives, and incorporate those perspectives and ideas into lessons. Similarly, teachers with a high Quality of Feedback score will have back-and-forth exchanges with students and clarify or further query student responses in an effort to expand student involvement and increase learning. These two dimensions may differentiate slightly (and similarly) from the other two Instructional Support dimensions due to their emphasis on student-centered instructional methods.

Given the variability of the factor structure across studies, further investigations of the internal structure of CLASS should systematically examine structural validity across grade levels. Specifically, researchers must test the fit of the bifactor and three factor models at each developmental level (pre-k, primary, intermediate) to determine which structure best represents CLASS dimensions and domains at each level. In addition, based on the Pakarinen study (2010) there also may be cultural differences in teaching practices that have implications for the structural validity of CLASS when used in a cross-cultural context. As such, future studies should examine the factor structure of CLASS across cultures to determine if there are variations in teaching practices and teacher–child interactions.

Structural validity evidence is essential, but not sufficient to determine the validity of assessment scores when used for a specific purpose. Future studies should continue to examine the relation between the various CLASS models (original and alternative) and children's academic and social-emotional outcomes. Linking the various models tested in the current study to children's functioning will provide additional, and critical, insight regarding the validity of CLASS scores based on different structures. Further, evidence for concurrent or predictive validity may be enhanced by linking CLASS scores with student outcomes using models that include mediating or moderating influences. For example, factors such as student motivation, engagement, perceptions of the classroom environment, and feelings of relatedness or self-efficacy may interact with CLASS domains to improve student academic and social-emotional outcomes (e.g., Martin & Rimm-Kaufman, 2015). In addition to looking at how CLASS relates to student outcomes, CLASS is also used as a professional development tool for teachers. Another way to explore the validity of CLASS is to continue examining the utility of this measure as a formative assessment that may advance or accelerate teachers' own professional development (e.g., Pianta, Mashburn, Downer, Hamre, & Justice, 2008).

There are two primary limitations to this study. First, the classroom sample (N = 141) is within the lower range of acceptable sample sizes for CFA given the number of parameters estimated in each model (MacCallum, Widaman, Zhang, & Hong, 1999; Kline, 2013). As such, the parameter estimates reported for the six models may be less reliable than those emerging from analyses with larger samples. Second, several of the models tested in this study—including the best-fitting model—included one or more correlated residuals. Although the correlation of residuals is statistically informative as it improves model fit and reveals the presence of additional shared variance among indicators, such modifications may simply take advantage of sample-specific variation to improve model fit and may not generalize. However, it is important to note that observed indicators will often share variance that is not related to the underlying latent construct, and this shared method variance may be important to explore when trying to better understand a structural model (Cole, Ciesla, & Steiger, 2007). Future research attempting to replicate the best-fitting CLASS structure from the current study should include the residual correlations to see if they generalize and use theory to guide the inclusion of any additional residual correlations.

Given the complex nature of teacher–child relationships and the many factors that potentially influence classroom quality, it is not surprising that these interactions can be difficult to measure. A primary conclusion of the large-scale Measures of Effective Teaching project (Kane & Staiger, 2012) was that instructional quality and teacher–child interactions are best understood through a multidimensional lens that includes observations, rating scales, and student achievement. Systematic observations, such as CLASS, are a primary method by which educational researchers and district administrators collect data on teachers. Given the widespread use of this method, it is essential to understand how aspects of teaching may change depending on contextual characteristics, as well as to continue to identify potential sources of error in the measurement of high-quality teaching constructs. The continued pursuit of more precise measurement of teacher effectiveness and teacher–child interactions will help to ensure that teaching practices are being evaluated appropriately for varying developmental levels and populations.

To date, the structural validity of CLASS has been evaluated across multiple studies in early childhood and elementary classrooms. Although stable findings emerged across the current and previous studies, particularly in relation to the presence of three latent domains, there is also some variability across structures at different grade levels with regard to the bifactor and three-factor models. Systematic examination of the factor structure across developmental levels would further enhance understanding of the most valid structure for interpretation of CLASS data in the prekindergarten, primary, and intermediate grades.

References

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin, 88,* 588–606. http://dx.doi.org/10.1037/0033-2909.88.3.588

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.

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Developmental Science, 12,* 140–153. http://dx.doi.org/10.1080/10888690802199418

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41,* 189–225. http://dx.doi.org/10.1207/s15327906mbr4102_5

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods, 12,* 381–398. http://dx.doi.org/10.1037/1082-989X.12.4.381

Curby, T. W., Grimm, K. J., & Pianta, R. C. (2010). Stability and change in early childhood classroom interactions during the first two hours of a day. *Early Childhood Research Quarterly, 25,* 373–384. http://dx.doi.org/10.1016/j.ecresq.2010.02.004

Curby, T. W., Rimm-Kaufman, S. E., & Abry, T. (2013). Do emotional support and classroom organization earlier in the year set the stage for higher quality instruction? *Journal of School Psychology, 51,* 557–569. http://dx.doi.org/10.1016/j.jsp.2013.06.001

Curby, T. W., Rimm-Kaufman, S. E., & Ponitz, C. C. (2009). Teacher–child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology, 101,* 912–925. http://dx.doi.org/10.1037/a0016647

Eccles, J. S. (1999). The development of children ages 6 to 14. *The Future of Children, 9,* 30–44. http://www.jstor.org/stable/1602703. http://dx.doi.org/10.2307/1602703

Hamre, B. K., Hatfield, B. E., Pianta, R. C., & Jamil, F. (2014). Evidence for general and domain specific elements of teacher-child interactions: Associations with preschool children's development. *Child Development, 85,* 1257–1274. http://dx.doi.org/10.1111/cdev.12184

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 early childhood and elementary classrooms.* New York, NY: Foundation for Child Development. Retrieved from http://fcd-us.org/resources/building-science-classrooms-application-class-framework-over-4000-us-early-childhood-and-e?destination=resources%2Fsearch%3Ftopic%3D0%26authors%3DHamre%26keywords%3D

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

Jones, S. M., Molano, A., Brown, J. L., & Aber, J. L. (2013). Reconceptualizing the CLASS framework in elementary schools: Domain specific links to teacher and child outcomes. In B. Hatfield (Chair), *Domain-general and Domain-specific Associations of the Classroom Assessment Scoring System to Children's Development from Preschool to Fifth Grade*. Paper presented at the Society for Research in Child Development, Seattle, WA.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains*. Policy and practice brief prepared for the Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kline, R. B. (2013). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Kyriacou, C. (2001). Teacher stress: Directions for future research. *Educational Review, 53,* 27–35. http://dx.doi.org/10.1080/00131910120033628

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., . . . Barbarin, O. (2007). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly, 22,* 3–17. http://dx.doi.org/10.1016/j.ecresq.2006.05.001

Määttä, K., & Uusiautti, S. (2012). Pedagogical authority and pedagogical love: Connected or Incompatible? *International Journal of Whole Schooling, 8,* 21–39.

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

Madill, R. (2014). How to teachers shape children's social development? A study of teachers' use of seating arrangements and responsive teaching. *Dissertation Abstracts.*

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43,* 304–316. http://dx.doi.org/10.3102/0013189X14545513

Martin, D. P., & Rimm-Kaufman, S. E. (2015). Teacher emotional support moderates the relation between math self-efficacy and engagement in fifth grade math classrooms. *Journal of School Psychology, 53,* 359–373. http://dx.doi.org/10.1016/j.jsp.2015.07.001

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79,* 732–749. http://dx.doi.org/10.1111/j.1467-8624.2008.01154.x

McCormick, M. P., & O'Connor, E. E. (2015). Teacher–child relationship quality and academic achievement in elementary school: Does gender matter? *Journal of Educational Psychology, 107,* 502–516. http://dx.doi.org/10.1037/a0037457

Muthén, L. K., & Muthén, B. O. (2008–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

National Center for Education Statistic. (2014). *Search for public schools*. Retrieved from https://nces.ed.gov/ccd/schoolsearch

Pakarinen, E., Lerkkanen, M., Poikkeus, A., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J. (2010). A validation of the Classroom Assessment Scoring System in Finnish kindergartens. *Early Education and Development, 21,* 95–124. http://dx.doi.org/10.1080/10409280902858764

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *The Classroom Assessment Scoring System Manual, K–3*. Baltimore, MD: Brookes Publishing Co.

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interaction in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 23,* 431–451. http://dx.doi.org/10.1016/j.ecresq.2008.02.001

Rimm-Kaufman, S. E., & Pianta, R. C. (2000). An ecological perspective on the transition to kindergarten: A theoretical framework to guide empirical research. *Journal of Applied Developmental Psychology, 21,* 491–511. http://dx.doi.org/10.1016/S0193-3973(00)00051-4

Rudasill, K. M., Reio, T. G., Jr., Stipanovic, N., & Taylor, J. E. (2010). A longitudinal study of student-teacher relationship quality, difficult temperament, and risky behavior from childhood to early adolescence. *Journal of School Psychology, 48,* 389–412. http://dx.doi.org/10.1016/j.jsp.2010.05.001

Sandilos, L. E., DiPerna, J. C., & The Family Life Project Key Investigators. (2014). Measuring quality in kindergarten classrooms: Structural analysis of the Classroom Assessment Scoring System, Kindergarten—Third Grade (CLASS K–3). *Early Education and Development, 25,* 894–914. http://dx.doi.org/10.1080/10409289.2014.883588

Schermelleh-Engel, K., Moosbrugger, H., & Muller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research, 8,* 23–74.

Whitaker, R. C., Becker, B. D., Herman, A. N., & Gooze, R. A. (2013). The physical and mental health of Head Start staff: The Pennsylvania Head Start staff wellness survey, 2012. *Preventing Chronic Disease, 10,* E181. http://dx.doi.org/10.5888/pcd10.130171

Zaslow, M., Martinez-Beck, I., Tout, K., & Halle, T. (2011). *Quality measurement in early childhood settings*. Baltimore, MD: Brookes.

Received October 1, 2015
Revision received February 21, 2016
Accepted February 27, 2016 ■