Evidence Regarding the Domains of the CLASS PreK in Head Start Classrooms

Rachel A. Gordon and Fang Peng

University of Illinois at Chicago

Early Childhood Research Quarterly, 53, 23-39 (Published March, 2020).

Rachel A. Gordon, Department of Sociology, University of Illinois at Chicago; Fang Peng, Department of Educational Psychology, University of Illinois at Chicago.

This paper was supported by grant R305A130118 from the Institute of Education Sciences and we gratefully acknowledge the entire project team. The results and statements do not necessarily reflect the views of our funders, and all errors and omissions are our own.

Correspondence concerning this article should be addressed to Rachel A. Gordon, PhD, Professor, Department of Sociology (MC 312), University of Illinois at Chicago, 1007 West Harrison Street, Chicago, IL 60607, E-mail: ragordon@uic.edu.

Keywords: classroom quality; academic achievement; social-emotional competency; factor analyses; bifactor model

Abstract

The standard scoring of the CLASS PreK produces three domain scores that are widely used in research, practice and policy. Despite these domains being based on developmental theory and research, limited empirical evidence exists for the three-domain structure as operationalized in the CLASS PreK. Using the 2009 and 2014 Head Start Family and Child Experiences Surveys (FACES), we estimated a series of exploratory and confirmatory bifactor and traditional factor analyses to produce evidence regarding this structure and possible alternatives in nationally representative samples of Head Start classrooms. Replicating and extending the small set of prior factor analytic studies, we found alternative factor structures fit equally well or better than the standard structure in both FACES 2009 and 2014 as well as problems with estimation and fit of a bifactor structure of general and specific factors proposed by the CLASS PreK developers. Across all domain structures, associations with children's academic and social-emotional gains during the Head Start year were uniformly small and generally non-significant. Our findings encourage future refinements of the CLASS PreK and continued development of new measures to better operationalize its conceptually-motivated domains.

Evidence Regarding the Domains of the CLASS PreK in Head Start Classroom

Over the last decade, the Classroom Assessment Scoring System (CLASS) rapidly became a dominant classroom observational system in research, practice, and policy. Head Start opted for a version of the CLASS designed for preschool classrooms (CLASS PreK; Pianta, La Paro, & Hamre, 2008) in its Designation Renewal System (DRS; Public Law 110-134) when aiming to ensure that all funded programs had sufficiently high quality to support program goals. The national Head Start Family and Child Experiences Survey (FACES) added the CLASS PreK to its core assessments in the late 2000s (Malone et al., 2013). And, a 2017 state scan found that the CLASS PreK was used in just over half of the state Quality Rating and Improvement Systems (QRIS) that incorporated observational tools (QRIS Compendium, 2017). The CLASS' standard scoring calculates values for three domains, and this standard scoring is regularly used in such research, practice, and policy applications. Despite these domains being based on developmental theory and research, limited empirical evidence exists for the three-domain structure as operationalized in the CLASS PreK. In a recent review of the use of the CLASS PreK in the Head Start DRS, Mashburn (2017) noted a particular need for studies demonstrating its domain structure in Head Start classrooms. Our paper offers such evidence, based on data from FACES 2009 and 2014. We also place these Head Start specific results in the context of the existing literature regarding the factor structure of the CLASS PreK and make recommendations for future research, practice, and policy based on the findings.

Prior Research on the Standard Domains of the CLASS PreK

The CLASS PreK emerged from a synthesis of developmental theory and research, emphasizing teacher-student (adult-child) interactions as the main driver of children's development and learning (Pianta et al., 2008). Based on this literature, the ten CLASS PreK items (referred to as "dimensions" by the scale developers) were organized into three domains—

Emotional Support, *Classroom Organization*, and *Instructional Support* (see Table 1 for a list of the ten item names by domain; Pianta et al., 2008). The manual describes Emotional Support as capturing "teachers' abilities to support social and emotional functioning in the classroom," Classroom Organization as encompassing a "broad array of classroom processes related to the organization and management of students' behavior, time, and attention," and Instructional Support as the ways in which teachers implement curricula (whatever their content) "in order to effectively support cognitive and language development" (Pianta et al., 2008, p. 3-5).

Empirically, this structure can be validated using factor analyses that: a) see if a three-factor structure fits better than other structures such as a single-factor structure; b) verify that the three factors are empirically distinct, with moderate-to-small intercorrelations, and, c) see whether the items load on the expected domains (i.e., as listed in Table 1). Few such factor analytic studies have been conducted, and those that do have not found strong evidence for the proposed structure. In their study of the CLASS PreK domains, Hamre and colleagues (2014, p. 1258) noted the three main limitations of existing factor analytic evidence as: (a) "less than ideal" absolute fit indices, despite relatively better fit of the three-domain structure than the one or two-domain structures, (b) "very high" correlations among the three domains, and (c) "small" effect sizes in associations of all three domains with children's developmental outcomes.

These conclusions apply to a small set of recently published factor analytic studies of the CLASS PreK that have examined its domain structure in the U.S. (multiple cities in Hamre et al., 2014; Boston in Weiland, Ulvestad, Sachs, & Yoshikawa 2013), Germany (Von Suchodoletz, Fäsche, Gunzenhauser, & Hamre, 2014), Chile (Leyva et al., 2015), Finland (Pakarinen et al., 2010) and Portugal (Cadima, Verschueren, Leal, & Guedes 2016). Consistent with the limitations in the literature summarized by Hamre and colleagues (2014), these studies have reported that the three-factor model fit better than the one- or two-factor models, although

absolute levels of fit typically fell below conventional cutoffs. When reported, associations with child outcomes were nonsignificant and/or small (in the Chilean study, .10 and below in linear models; in the Boston study, no significant linear associations and associations up to .20 in non-linear models, although sometimes in unexpected directions). Correlations among the factors were also moderate to high, in the studies reporting them (from .63 to .76 in Germany; .69 to .86 in Boston; above .90 in Finland). Also reflective of shared variance among factors, the Chilean and Finnish studies reported needing to release residual correlations among items in order to raise fit to acceptable levels.

Although most studies focused on the three domains reflected in the standard CLASS scoring, some identified alternative structures. In the Portugese sample, Cadima and colleagues (2016) found a two-factor solution fit better than the standard three-factor solution, combining together the Emotional Support and Classroom Organization domains. The correlation between the two resulting factors remained sizable (r = .67), however, and associations with children's self-regulation scores were often non-significant and small (non-significant in the full sample; standardized coefficients of approximately .20 for the subgroup of children who started school with lower self-regulation among the full sample of 206 children). The Finnish study also adjusted the three-factor solution by excluding one of the CLASS PreK items (negative climate) due to its high correlation with items from two domains (Emotional Support and Classroom Organization).

Prior Research on Alternative General/Specific Domains Proposed by the CLASS PreK Authors

In a publication co-authored by two of the CLASS PreK developers (Hamre, Hatfield, Pianta, & Jamil, 2014), an alternative domain structure was proposed that might address the limitations evident in prior studies. This structure had a general domain that included all ten

CLASS PreK items—referred to by the authors as *Responsive Teaching*. The authors described this as a "general dyadic systems-level property of teacher-child interactions...hypothesized to foster children's development across all domains (social-emotional, behavioral, and cognitive outcomes)" (Hamre et al., 2014, p. 1258). This general domain could absorb the variance shared by all items, allowing additional variance shared only by certain items to be captured by specific domains reflective of "unique elements of teachers' interactions with children...hypothesized to show differential associations to outcomes in social and cognitive domains" (Hamre et al., 2014, p. 1258). The authors hypothesized three specific factors mirroring the three standard domains but relabeled *Motivational Supports*, *Proactive Management and Routines,* and *Cognitive Facilitation*. These specific classroom quality domains were expected to align with and uniquely support three respective domains of children's development: a) positive social relationships, b) stronger executive functioning, and c) better academic achievement (especially language and literacy skills; Hamre et al., 2014, p. 1261).

Such a structure, that includes a general domain encompassing all items as well as specific domains encompassing only certain items is referred to in the factor analytic literature as a *bifactor* structure (Gibbons & Hedeker, 1992). The bifactor structure has recently gained popularity in numerous fields due to its potential for addressing high empirical correlations among conceptually distinct constructs (Lahey et al., 2017; Reise, 2012; see Colwell, Gordon, Fujimoto, Kaestner, & Korenman, 2013 and Hindman, Pendergast & Gooze, 2016 for recent applications of the bifactor model to other observational preschool quality measures). The bifactor model distinguishes the general and specific factors in such a way that the resulting factors are uncorrelated. However, the model rests on assumptions that are not always well met in the data. Problems during analysis such as residual variances being outside of plausible values (i.e., negative) and models failing to meet convergence criteria are signals that the model does

not well fit the data (Eid, Geiser, Koch, & Heene, 2017).

Such convergence problems were reported by Hamre and colleagues (2014, p. 1265) when they initially tried to estimate the proposed general plus three-specific-factor bifactor structure. As a result, they modified the originally hypothesized structure by reducing from three to two domain-specific factors (combining together *Motivational Supports* with *Proactive Management and Routines*, and using the latter name for the combined factor). The authors reported that the model with this structure converged and met absolute fit targets for some criteria, although three items did not load significantly on a domain-specific factor (teacher sensitivity, regard for student perspectives, and instructional learning formats). Associations with child outcomes were also small (standardized coefficients of .08 or less). The results followed the expected domain-specific pattern in the sense that Cognitive Facilitation associated significantly with language and literacy outcomes (standardized coefficients of .05 and .06) as did Positive Management and Routines with a measure of executive functioning (the Pencil Tap; standardized coefficient of .07). However, these associations were also evident with the standard domains (.08 and .09 for Instructional Support associating with language and literacy outcomes; .14 for Classroom Organization associating with executive functioning). Altogether, the empirical evidence for the bifactor structure was limited, and the resulting factor scores did not demonstrate stronger associations with children's development than did the standard domain scores.

Replication of the bifactor structure with the CLASS PreK. To our knowledge, a replication of the Hamre and colleagues' (2014) bifactor structure with the CLASS PreK in a U.S. sample has not yet been published, although two studies have done so using Chinese (Hu, Fan, Gu, & Yang, 2016) and German (Bihler, Agache, Kohl, Willard, and Leyendecker, 2018) classrooms. The first study included 180 classrooms from three municipalities in China's

Guangdong province, finding that a bifactor model matching the Hamre et al. two-specific-domain structure had nearly equivalent fit to the standard three-domain model. The absolute levels of fit were below conventional cutoffs, however, and one item (regard for student perspectives) had a zero loading on its domain-specific factor. Bihler and colleagues (2018) studied 177 classrooms in the large German state of North Rhine-Westphalia. The authors found that the Hamre et al. two-specific-domain bifactor structure fit less well than the standard three-domain structure, although they adjusted the standard structure by freeing a cross-loading (allowing the language modeling item to load on the Emotional Support domain) and including two correlated errors (for positive climate with both quality of feedback and behavior management). The authors also reported a negative residual variance for behavior management that they fixed at zero, suggesting problems with model identification.

These two studies replicated the limited evidence for the bifactor structure being preferred over the standard domains offered by the Hamre et al. (2014) study, in the sense that they reported problems with model fit and with low loadings for some items. For both bifactor studies, the authors also noted possible cross-country normative and regulatory differences that may limit generalization to the U.S. For instance, Bihler and colleagues indicated that German preschool teachers received highly standardized training that emphasized free-play over group-based structured activities and that classrooms often mixed age groups. In the studied classrooms, about one-quarter of children fell outside of the recommended age range for the CLASS PreK (11% younger than age 3; 12% older than age 5). The Chinese classrooms, in contrast, were age segregated, with equal numbers of studied classrooms serving children in three age groups (3-, 4- and 5-year olds). However, class sizes were larger than in the German sample (averaging about 32 versus 21 children) and emphasized whole-group instruction.

Replication of the bifactor structure with the CLASS K-3 and CLASS-S. Given the

few factor analytic studies of CLASS PreK, several other studies applying the bifactor model to versions of the CLASS designed for elementary and secondary classrooms are informative. Three U.S. based studies examined the CLASS for kindergarten to third grade (CLASS K-3), which has the same domains and items as the CLASS PreK (Madill, Gest, & Rodkin, 2013; Sandilos, DiPerna, & the Family Life Project Key Investigators, 2014; Sandilos, Shervey, DiPerna, Lei, & Cheng, 2017; see also Longobardi, Pasta, Marengo, Prino, & Settanni, 2018, for similar results in an Italian sample using the CLASS K-3). One U.S. based study examined the CLASS for secondary grades (CLASS-S), which has the same domains but somewhat different items than the CLASS PreK (Hafen et a., 2015). Authors of these studies reported that a bifactor model similar to the structure identified by Hamre et al. (2014) was sometimes best fitting, although some loadings varied. Some of these studies also reported alternative traditional factor structures that fit as well or better, with revised item loadings and residual correlations different from the standard structure. These results generally differed from the standard CLASS scoring in terms of their placement of items from the standard Emotional Support and Classroom Organization domains or residual correlations involving these items.

Patterns of Results Across Prior Studies

Across the published studies, two patterns of results emerged that suggested possible alternative structures for the CLASS PreK. The first involved the number of domains and their corresponding items. Regarding the number of domains, the Instructional Support domain was consistently evident as a separate factor, but items from the Emotional Support and Classroom Organization domains were often combined. Regarding corresponding items, the Emotional Support and Classroom Organization domains were sometimes each evident, although not all of their corresponding items loaded highly, items cross-loaded on a domain different from the standard scoring, or item residuals correlated across the standard domains (Cadima et al., 2016;

Hafen et al., 2015; Hamre et al., 2014; Hu et al., 2016; Madill et al., 2013; Sandilos et al., 2014, 2017). Cadima and colleagues (2016), for instance, found that the Emotional Support and Classroom Organization items combined in a single factor in their traditional factor analysis. These results were consistent with Hamre et al. placing these first seven items together in one domain-specific factor, although teacher sensitivity, regard for student perspectives, and instructional learning formats loaded at zero or negatively on a domain-specific factor.

The second general pattern evident in prior studies relates to problems with convergence. As noted, Hamre et al. (2014) found that their first attempted bifactor structure—specifying three domain-specific factors corresponding to the three standard CLASS domains—failed to converge, Pakarinen and colleagues (2010) reported a negative residual variance for the language modeling item, Bihler and colleagues (2018) reported negative residual variances for the language modeling and behavior management items, and Sandilos and colleagues (2017) reported a negative residual variance for the behavior management item. Potentially related to these convergence problems is skewness in CLASS item scores, whereby the emotional support item scores are often concentrated at the high end and the instructional support item scores concentrated at the low end. Prior factor analytic studies have generally not discussed detailed screening for outlying and influential cases, although Sandilos and colleagues (2014) reported that they identified substantial skewness for the negative climate item along with nine extreme outlying cases. These authors then estimated a model that used the logged form of the negative climate item and that excluded the extreme cases. As noted above, Pakarinen and colleagues (2010) excluded the negative climate item altogether from their analyses.

Given these two patterns of results, we anticipated possible configurations of the first seven items that varied from the standard CLASS scoring. As Hamre et al. (2014) posited, these items may tap into the broad domain of responsive teaching, more specific domains of

motivational supports and proactive management and routines, or some combination of these constructs. In our study, exploratory analyses of both traditional and bifactor structures allow us to see whether a consistent alternative structure is evident in the FACES 2009 and FACES 2014 samples. In examining these and other structures, we also expected that attention to potential item skewness and convergence problems would be important. Screening for outlying observations and comparing results when including and excluding or when transforming extreme values informs us as to the sensitivity of results to such cases (Christensen, Freese, & Miguel, 2019). Possibly, the varying results across prior studies in part reflects such outlying cases having considerable influence, which may be particularly evident given the relatively modest size of many samples. Screening for outlying observations may also reduce problems with convergence and better support recovery of conceptually expected factor structures.

Current Study

In the current study, we replicated and extended recent factor analytic research based on CLASS PreK scores. We focused on Head Start classrooms, recognizing the program's broad importance in the field of early care and education and the specific need to consider whether the use of the three standard CLASS PreK domains in the Head Start DRS is empirically supported. Our research questions were: (a) What alternative factor structures for the CLASS PreK are suggested by exploratory factor analyses and item screening? (b) What is the best-fitting domain structure when comparing confirmatory models? (c) How intercorrelated are the identified domains? (d) How do standard and factor scores relate to gains in children's academic and social-emotional scores? We considered one-, two-, and three-domain traditional and bifactor structures.

Method

Sample

We used data from the 2009 and 2014 Head Start Family and Child Experiences Surveys (FACES). The FACES studies provide nationally representative samples of Head Start children and the classrooms that they attended. Importantly for our study, the FACES surveys followed children from fall enrollment to spring completion of their first Head Start year, also observing their attended classrooms in the spring.

Data were collected by the policy research organization Mathematica under contract from the Office of Planning, Research, and Evaluation in the Administration for Children and Families of the U.S. Department of Health and Human Services (Klein et al., 2017; Malone et al., 2013). To achieve a nationally representative sample with geographic locations that were feasible and cost-effective for deploying field interviewers (Heeringa, West & Berglund, 2010), survey statisticians from Mathematica developed multi-stage stratified and clustered designs that first randomly sampled Head Start programs from administrative lists provided by the federal Head Start agency and then sampled centers from programs, classrooms from centers, and children from classrooms. Mathematica created and released variables that we used to adjust for design features in order to produce accurate nationally-representative estimates and associated standard errors (clustering in primary sampling units [PSUs], membership in strata, and sampling weights to adjust for initial sampling probabilities, attrition over time, and nonresponse to specific instruments; Heeringa, West & Berglund, 2010; Klein et al., 2017; Malone et al., 2013).

We implemented several decision rules to define the analysis sample for our study. These were separately implemented for our two types of analyses: a) factor analyses using classroom-level data; and, b) regression analyses using both classroom- and child-level data.

For the *factor analysis samples*, we focused on the classrooms that had CLASS PreK observations (n = 370 in spring 2010; n = 641 in spring 2015). In FACES 2009, these observed classrooms reflected a random subsample due to budget constraints. CLASS PreK observations

were conducted at or near the end of the Head Start year (March through May in spring 2010; March through June in spring 2015). The completion rates of eligible classrooms were high (98% in both years; Klein et al., 2017, p. 154; Malone et al., 2013, p. 149). Mathematica created sampling weights specifically for analyses of the classrooms that had classroom observations, adjusting for initial sampling probabilities, attrition over time, and nonresponse to specific instruments (Klein et al., 2017, p. 188; Malone et al., 2013, p. 150). We apply these classroom-level weights in our factor analyses.

For the *regression analysis samples* we focused on children who were followed from fall to spring. FACES 2009 (and earlier FACES studies) focused on newly enrolled children—those beginning Head Start for the first time, excluding those returning for a second Head Start year. For comparability, we focused on the newly enrolled children in FACES 2014. Completion rates for child assessments were high in the fall (94% in 2009; 95% in 2014), and lower in spring reflecting attrition that included children leaving Head Start before the end of the year (85% in both 2010 and 2015; Malone et al., 2013, p. 119-120; Klein et al., 2017, p. 152). In FACES 2009, our analysis sample contained n = 2,381 children and n = 369 classrooms (one observed classroom had no child-level data); In FACES 2014, the sample had n = 974 children and n = 193 classrooms (some classrooms had no children with assessment data by design). Mathematica created child-level sampling weights specifically for longitudinal analyses like ours that used children's fall and spring assessment data, as well as their classroom's observation data, and adjusted for initial sampling probabilities, attrition over time, and nonresponse to specific instruments (Malone et al., 2013, p. 151; Klein et al., 2017, p. 189). In our multi-level regression models (described below), we used these weights at the child level and we used the classroom-level weights at the classroom level.

Measures

CLASS PreK. Observer ratings were based on the CLASS PreK's ten items (referred to as dimensions in the manual) assessing the quality of classroom interactional processes in its three broad domains (see again Table 1; Pianta et al., 2008). Observers used narrative descriptions of the *low* (1, 2), *middle* (3, 4, 5) and *high* (6, 7) range of each item when assigning scores. The CLASS PreK manual indicates that: "Because of the highly inferential nature of the CLASS, scores should never be given without referring to the manual" (Pianta et al., 2008, p. 17). Observers are certified to score the CLASS PreK by completing official training from the company Teachstone, founded by the CLASS PreK developers. Training includes multi-day in-person training by a Teachstone-certified trainer as well as Teachstone certification based on meeting criteria for within-one agreement with master ratings of Teachstone-supplied videos.

All classroom observers for FACES 2009 and 2014 were certified through Teachstone's certification process (including using Teachstone videos and following Teachstone procedures to assess reliability with Teachstone-developed master codes for those videos; Mathematica, personal communication, August 5, 2019). Beyond Teachstone certification, Mathematica also required that FACES observers demonstrated reliability with Mathematica-designated gold standard observers during a live classroom observation. Gold standard observers were Mathematic staff who were themselves certified CLASS observers and experienced in conducting classroom observations. Mathematica required the field staff observer's score to be within one point of the gold standard observer at least 80% of the time in order to be certified as a FACES observer (Mathematica, personal communication, August 5, 2019). Twenty observers were certified for observations in spring 2010 as were 48 observers in spring 2015 (Malone et al., 2013, p. 101; Klein et al., 2017, p. 130).

Mathematica also monitored reliability twice during each field period (Klein et al., 2017, p. 149; Malone et al., 2013, p. 118). To do so, a gold standard observer joined a FACES-certified

field staff observer to score the same classroom. The same criteria as the original live observations was used to monitor for drift. Within-one percentage agreement averaged 95% in spring 2010 (Malone et al., 2013, p. 118) and 90% in spring 2015 (Klein et al., 2017, p. 150).

Following standard CLASS procedures, observers generally scored each item during four 20-minute cycles (in some classrooms with shorter total observation periods, only three cycles were scored; 33% had three cycles in spring 2010, 18% had three cycles in spring 2015; Klein et al., 2017, p. 241; Malone et al., 2013, p. 206). Each item's categorical ratings were averaged across these cycles. Internal consistency values for spring 2010 and spring 2015, respectively, were .82 and .80 for Emotional Support, .77 and .82 for Classroom Organization, and .87 and .89 for Instructional Support (Malone et al., 2013, p. 204-205; Klein et al., 2017, p. 250). The resulting quasi-continuous values were available in the FACES data releases. We factor analyzed these values, consistent with prior studies that relied on the standard CLASS scoring and also used such across-cycle quasi-continuous average scores. Like other factor analytic studies of the CLASS, we reverse coded scores for the only negatively oriented item (negative climate). As a result, higher scores represented higher quality on each item.

Child assessments. The same child assessments were used in both FACES 2009 and 2014 in most cases, with exceptions noted below. Within samples, children were assessed on the same instruments in the fall and the spring. Altogether we analyzed four academic assessments in both samples, administered in English or Spanish, as well as six social-emotional assessments for FACES 2009 and four social-emotional assessments for FACES 2014.

Field staff were comprehensively trained for the fall child assessments through classroom discussion, paired practice and certification. Particular attention was paid to mastery of screening children for language ability and interacting with children of varying developmental level, language background, and disability status. Field staff were certified by observation of their

administering the assessments to a 3- or 4-year-old child. Trained certifiers used a standard certification form that rated technical accuracy as well as rapport and fluidity. A score of 90% (of 465 points in 2009 and of 521 points in 2014) was required. Bilingual staff attended additional training and were certified in both the English and Spanish instruments. Refresher training and recertification (using the same criteria) was required for the spring child assessments (Malone et al., 2013, p. 96-98, 102; Klein et al., 2017, p. 123-127).

Academic outcomes. Children were routed into English or Spanish language academic assessments based on their home language and their performance on the Simon Says and Art Show English language screening measures from the Preschool Language Assessment Survey (Duncan & DeAvila 1998). Sample sizes in each group are provided below.

In both samples, one academic outcome was the *Peabody Picture Vocabulary Test, Fourth Edition* (*PPVT-4*; Dunn, Dunn, & Dunn, 2006) which assessed children's receptive (hearing) vocabulary. The assessor presented a series of words to the children, each accompanied by four pictures. The children were asked to say the number or point to one of the four pictures corresponding to each word. The PPVT has scale-developer reported test-retest reliability ranging from .92 to .96 (Dunn et al., 2006). FACES investigators reported internal consistency of .97 in fall 2009 and .95 in spring 2010 (Malone et al., 2013, p. 177) and of .97 in fall 2014 and spring 2015 (Klein et al., 2017, p. 219). We focused on standard scores based on the scale developers' nationally representative norming sample of children and adults from across the U.S. In FACES 2009, Spanish-speaking children completed the *Vocabulario de Imagines Peabody* (*TVIP*; Dunn, Padilla, Lugo, & Dunn, 1986). In FACES 2014, the Spanish Edition of the *Receptive One-Word Picture Vocabulary Test–4*: (ROWPVT) was used to assess children's receptive vocabulary in Spanish (Martin & Brownell, 2012). FACES reported internal consistency of .93 (fall) and .94 (spring) for the TVIP (Malone et al., 2013, p. 177) and .96 in

both fall and spring for the ROWPVT (Klein et al., 2017, p. 220).

In both samples, children also completed three batteries of the English language *Woodcock-Johnson III Tests of Achievement–Third Edition* (Woodcock, McGrew, & Mather, 2001) or the Spanish language *Batería III Woodcock-Muñoz* (Woodcock, Muñoz-Sandoval, McGrew, Mather, & Schrank, 2004). The *Applied Problems* subtest measured children's ability to perform simple counting, addition, or subtraction operations. The *Letter-Word Identification* subtest measured children's skill in identifying letters and words printed on test book pages. The *Spelling* subtest assessed pre-writing and writing skills, such as drawing lines, copying letters, and writing words. Internal consistency reliability as reported by FACES investigators was moderate to high for FACES 2009 (English: .85, .87, .79 in fall for Letter Word, Applied Problems, and Spelling respectively; .88, .89, .83 in spring; Spanish: .67, .84, .66 in fall; .85, .87, .73 in spring; Malone et al., 2013, p. 178-179) and FACES 2014 (English: .87, .89, .81 in fall for Letter Word, Applied Problems, and Spelling respectively; .90, .88, .84 in spring; Spanish: .80, .84, .66 in fall; .82, .87, .71 in spring; Klein et al., 2017, p. 220-222).

Social-emotional outcomes. For social-emotional development, in FACES 2009, one set of measures relied upon both *teacher and parent reports* of children's *social skills* (cooperative, empathic, and helpful behaviors) and *behavior problems* (aggressive, hyperactive, anxious, and withdrawn behaviors). In FACES 2014, only the teacher reports were used. In both FACES samples, teachers responded to 12 social skills items that FACES investigators drew from the Social Skills Rating System (Elliott, Gresham, Freeman, & McCloskey, 1988) and the Personal Maturity Scale (Entwisle, Alexander, Cadigan, & Pallis, 1987). Fourteen teacher-reported behavior problems items came from the Behavior Problems Index (Zill, 1990) and the Personal Maturity Scale (Entwisle et al., 1987). Teachers rated both sets of items on a 3-point scale: 0 (*Never*), 1 (*Sometimes*), and 2 (*Very Often*). FACES investigators reported good internal

consistency values for these scales (FACES 2009: .89 in both fall and spring for social skills; .88 and .87 for problem behaviors; Malone et al., 2013, p. 183; FACES 2014: .89 in fall and .91 in spring for social skills; .86 and .87 for problem behaviors; Klein et al., 2017, p. 226). In FACES 2009, parents rated sets of 8 social skills items and 12 behavior problems items drawn from similar sources as the teacher items (Malone et al., 2013, p. 58-59). Parents reported about their children's behavior during the past month using a 3-point scale ranging from 0 (*Not true*) to 2 (*Very true or often true*). FACES investigators reported moderate internal consistency values for the parent scales (.68 and .69 in fall and spring for social skills; .72 and .73 for problem behaviors; Malone et al., 2013, p. 184). For both teachers' and parents' reports, we used average scores of each set of items in our models.

Examiners also rated children's behavior during the administration of the direct assessments using the academic/social subscale from the *Leiter International Performance Scale, Revised* (Leiter-R; Roid & Miller, 1997). Examiners rated 27 items encompassing children's attention, impulse control, activity, and sociability on a four-point scale: 0 (*rarely/never*), 1 (*sometimes*), 2 (*often*), and 3 (*usually/always*). We used standard scores in our models. FACES investigators reported internal consistency reliabilities of .90 in both fall and spring for these items in FACES 2009 (Malone et al., 2013, p. 184) and of .91 in fall and .90 in spring in FACES 2014 (Klein et al., 2017, p. 226).

The *Pencil Tapping* task (Blair 2002; Diamond & Taylor, 1996; Smith-Donald et al., 2007) directly assessed inhibitory control, working memory, and attention, by requiring the child do the opposite of what the assessor said (e.g., tap one time when the assessor said to tap two times; tap two times when the assessor said to tap one time). We analyzed the percentage of 16 tasks that the child completed correctly. FACES administered the task only to children who were 4-years or older, and reported internal consistency of .88 in fall 2009, .86 in spring 2010, .94 in

fall 2014, and .90 in spring 2015 (Klein et al., 2017, p. 223; Malone et al., 2013, p. 180).

Controls. We controlled for potential confounds that might be associated both with classroom quality and child development. One important control was the fall child assessment score corresponding to each spring child assessment. Other covariates included the child's race-ethnicity (Hispanic; non-Hispanic White, Black and Other), binary sex-gender (male, female), whether English was the child's assessment language, whether the family's income fell below 200% of the federal poverty line for their family size, whether the child had a special need or disability (as reported by the parent in 2009 and by the teacher in 2014), whether the mother's education level was high school or less, whether the child was in the 3- versus 4-year old age cohort in the fall, the child's age at the spring child assessment in months, and the time interval between the fall and spring child assessments (in months). Appendix A16 to A20 provides sensitivity analyses in which we added classroom-level covariates, including number and ages of children and the teacher's education and years of experience. These results led to the same conclusions as those presented in Tables 4-6 (just 14 of 240, or 6%, of coefficients were statistically significant, and, the coefficients were consistently small in size and sometimes of opposite sign than conceptually predicted).

Analysis Plan

Item screening. We began by presenting univariate and bivariate statistics and graphs of the CLASS items, applying the classroom-level sampling weights. In addition to ranges, means, and standard deviations, we reported skewness and kurtosis statistics based on the central moments of the item distributions, with values above one and two respectively considered extreme values (Bishara & Hittner, 2017). Box plots of the items complemented these statistics, visually characterizing the distributions. We identified multivariate outliers using Mahalanobis distance with a recommended cutoff of 29 based on the Chi-square distribution for our 10 items

(Riani, Atkinson, & Cerioli, 2009). We also presented item inter-correlations, again weighted by the classroom-level sampling weights.

Exploratory factor analyses. We estimated traditional exploratory factor analytic models with Geomin oblique rotation, and bifactor exploratory models with both Geomin orthogonal and oblique rotation, using a maximum likelihood estimator in Mplus Version 7.4 (Muthén & Muthén, 2012). In the exploratory context, all items were allowed to load on all factors. Therefore, fit indices were identical for the traditional and bifactor models, with the loadings reflecting the different rotations. In other words, a traditional exploratory solution with *c* factors corresponded to a bifactor exploratory solution with one general and *c-1* domain-specific factors.

Confirmatory factor analyses. For the confirmatory factor analytic models, we used maximum likelihood estimation in Mplus Version 7.4 (Muthén & Muthén, 2012), and identified the models by fixing the factor variances at one.

We estimated four traditional (non bifactor) confirmatory factor analytic models. Each of these models allowed an item to load on one and only one factor. The first (*Model 1a*) was the scale developers' original three-domain structure (which we refer to as the "*standard three-factor traditional*" structure). The second (*Model 1b*) was a three-factor structure identified in our exploratory models and apparent in the loading pattern of the Hamre et al. (2014) bifactor structure (which we refer to as the "*alternative three-factor traditional*" structure). The third (*Model 1c*) was a two-factor solution, in this case reflecting the alternative structure identified by Cadima and colleagues (2016) and the two domain-specific factors identified by Hamre and colleagues (2014), both of which combined the items from the Emotional Support and Classroom Organization domains (which we refer to as the "*alternative two-factor traditional*" or "*combine ES and CO*" structure). The fourth (*Model 1d*) was a "*one-factor traditional*" structure having all

items load on a single factor.

We also estimated five confirmatory bifactor models. These bifactor models allowed each item to load on a general factor and on one domain-specific factor as well as restricting all factors to be orthogonal (Gibbons et al., 2007; Reise, 2012). A first pair of models each had three domain-specific factors, one reflecting the standard CLASS domains (*Model 2a*, "*standard three-specific bifactor"*) and the alternative structure we identified (*Model 2b, "alternative three-specific bifactor"*). The second pair of models each had two domain-specific factors. These were the specifications that Hamre and colleagues (2014) found converged, the first (*Model 2c,* which we refer to as the *"Hamre et al. two-specific bifactor"* or "*combine ES and CO bifactor*") allowed each item to load on both the general factor and one of the domain-specific factors, and, the second (*Model 2d*, *Hamre et al. adjusted two-specific bifactor* or "*combine ES and CO bifactor adjusted*") restricted three items to load only on the general factor and not a specific factor (teacher sensitivity, regard for student perspectives, and instructional learning formats) due to Hamre et al.'s initial pattern of loadings. The final bifactor model had one domain-specific factor comprised of the Instructional Support items (*Model 2e, "one-specific bifactor"* or "*IS only*"), allowing for a general factor plus the domain-specific Instructional Support factor which was most consistently evident in prior studies.

Fit indices and loading patterns. Consistent with prior studies and following Brown (2015), we used the following criteria to indicate good fit of the factor analytic models: (a) comparative fit index (CFI) around .95 or above, (b) nonnormed fit index (NNFI) close to or above .95, and (c) the root-mean-square error of approximation (RMSEA) around .08 or below. Criteria for choosing among models based on these indices are not well established, although Cheung and Rensvold (2002) suggested that a difference of .01 or more for the CFI may be meaningful. We also reported the exact ΔCFI value between best-fitting models, so readers can

consider alternative cutoffs (Kang, McNeish, & Hancock, 2016). We reported exact factor loadings and looked for values ≥ .40 to indicate meaningful relationships to the factor. We checked for a *simple structure* in the exploratory context in which each item had a loading ≥ .40 on one factor and < .40 on all other factors (Brown, 2015).

Raw and factor scores. For traditional factor structures, we followed CLASS scoring to create *raw scores* by averaging item scores within the three standard domains and within alternative domains. For bifactor structures, we estimated *factor scores* using the regression method in Mplus (Skrondal & Laake, 2001).

Regression models. We conducted mixed models regressing children's spring outcome scores on their classrooms' spring CLASS scores, adjusting for the covariates described above (including the child's fall score on the instrument corresponding to the outcome). The models included random intercepts at the classroom level and applied the child sampling weights at the child level and the classroom sampling weights at the classroom level, using the size approach to scaling the multi-level weights (Rabe-Hesketh & Skrondal, 2006). Coefficients (and their standard errors) were standardized. For each outcome, we conducted a series of five regression models with different sets of CLASS scores as focal predictors. The first four sets used raw average scores: (a) for the three standard domains, (b) for the three alternative domains, (c) for the two domains (combining Emotional Support and Classroom Organization), and (d) a total raw score based on all ten items. The fifth set used the factor scores from the Hamre et al. two-domain-specific bifactor structure (combining Emotional Support and Classroom Organization).

We considered both the significance and size of associations between CLASS scores and child outcomes. For significance, we used a conventional *p*-value of .05, while also providing standard errors to denote precision of estimation. For size, we used standardized coefficients, which provide the expected standard deviation change in the outcome for a one standard

deviation increase in the CLASS PreK score. Because the covariates included the child's corresponding fall score on the outcome, these coefficients can be interpreted as gains. For instance, a coefficient of .08 would mean that children attending classrooms that differed by one standard deviation on the CLASS PreK had outcome gains that differed by just eight-hundredths of a standard deviation.

For these regression models, we used multiple imputation with chained equations to address item-level missing data (see Appendix A1-A4 for item-level missingness; most items had less than 5% missing cases with missingness being higher for the cognitive assessments as well as parent-reported household income, maternal education, child disability status, and behavior problems). For imputation of the English and Spanish language academic assessments we separated the children into two groups: (a) those whose home language was Spanish and were assessed in Spanish either in fall or spring and (b) those whose home language was English or another language other than Spanish and were assessed in English in both fall and spring. The majority of children (83% in 2009 and 86% in 2014) were in the English assessment group. Most children in the Spanish assessment group were assessed in Spanish at both time points (48% in 2009; 60% in 2014). Nearly all who had Spanish assessments at only one time point switched to English in spring; the exception was one child in 2014 who was assessed in Spanish in spring but not fall. Less than 1% of children in 2009 and 2014 spoke a language other than English or Spanish. We used conditional specifications to impute the appropriate child assessment given the child's group. We also used conditional imputation for the Pencil Tapping measure. Because this measure was completed only by children who were at least 4 years old, we imputed and analyzed only scores for the 4-year-old cohort who would have been age-eligible for both the fall and spring Pencil Tapping assessments.

Results

Description of Samples

The children in our analysis samples were balanced by gender (50% female in both 2009 and 2014) and diverse racial-ethnically (36% Hispanic, 22% non-Hispanic White, 35% non-Hispanic Black, 8% non-Hispanic Other in 2009, 38%, 31%, 23%, and 8%, respectively, in 2014; see again Appendix A3). As expected, based on Head Start enrollment guidelines, the majority of families (93% in 2009; 92% in 2014) had incomes below 200% of the federal poverty line. Sixty-nine percent of mothers in 2009, and 59% in 2014, had attained an education at the high school level or below. Mothers reported that 4% of children had a special need or disability that had been identified by a doctor or health professional in 2009. In 2014, teachers reported that 18% of children had special needs or disabilities (some including an Individualized Education Plan [IEP]). Somewhat more than half of children (58%) were in the 3-year-old cohort at fall 2009 enrollment; in 2014 the percentage was 61%. With an average interval of 5.5 to 6 months between the fall and spring child assessments, children averaged 53 months of age by the spring child assessment in both samples.

Research Question 1: What Alternative Structures Are Suggested by Item Screening and Exploratory Factor Analyses?

We begin with results of item screening and exploratory factor analyses. To conserve space, many results are detailed in referenced appendices available in online supplementary materials and summarized in the text.

Univariate item screening. Table 1 provides the CLASS items' univariate descriptive statistics and intercorrelations. Figure 1 shows the distributions graphically. We see in both Table 1 and Figure 1 the truncation and skewness of item scores also evident in prior studies. In these nationally representative samples of Head Start classrooms, the means listed in Table 1 revealed that most classrooms scored in the mid-to-high range on the first seven items from the

Emotional Support and Classroom Organization domains (*M*s = 4.00 to 6.75 in 2009; 4.25 to 6.75 in 2014) and in the low range on the final three items from the Instructional Support domain (*M*s = 2.12 to 2.45 in 2009; 2.26 to 2.55 in 2014). The boxplots in Figure 1 similarly showed the concentration of the distributions at or above four for the first seven items and at or below three for the final three items, although for each item some values appeared potentially outlying. In both samples, calculated values for skewness and kurtosis were extreme for the negative climate item; the skewness value for concept development was also just over the cutoff of |1| in 2014 (see again Table 1). Following Sandilos and colleagues (2014) we examined the sensitivity of our confirmatory factor analysis results when we used the natural log to reduce the skew of the negative climate item. Following Pakarinen and colleagues (2010) we also estimated models that excluded the negative climate item. Because these results revealed a similar pattern of fit and loadings, and to match the more common usage, we presented the results including nonlogged negative climate in the manuscript; we provided results based on the logged version and excluding the item in Appendix A5 to A8.

Turning to the inter-item correlations, we see in Table 1 that the values are small/moderate to high (*r* =.19 to .76 in 2009; -.03 to .82 in 2014), although the domain structure is not sharply evident. We used grey shading to highlight the within-domain correlations consistent with the standard CLASS scoring, and we bolded correlations above .50. These markings made salient that the correlations among the Instructional Support items were sizable (*r* =.61 to .76 in 2009; .62 to .82 in 2014) and larger than these items' cross-domain correlations with the other seven items (*r* =.19 to .49 in 2009; -.03 to .31 in 2014). In contrast, the within-domain correlations for the Emotional Support and Classroom Organization items tended to be smaller (*r* =.37 to .73 in 2009; .19 to .74 in 2014) and these items' cross-domain correlations tended to be larger (*r* =.21 to .62; .15 to .74 in 2014). Although the pattern of correlations was

broadly similar in 2009 and 2014, one notable difference was the somewhat lower cross-domain correlations of the Emotional Support and Classroom Organization items with the Instructional Support items in 2014, which stood out especially for the negative climate item, as well as the generally higher correlations within and between the Emotional Support and Classroom Organization items in 2014, again with the general exception of negative climate. Altogether, this pattern of correlations suggested a stronger signal in these samples of the Instructional Support than the Emotional Support and Classroom Organization domains.

Multivariate item screening. Our screening for potentially outlying cases identified five in 2009 and sixteen in 2014 with Mahalanobis distance values above the cutoff (data not shown in tables). Although these cases had the potential to influence results, we found that the fit and loadings of the factor models were similar when these classrooms were excluded. Given this pattern of results, we retained all classrooms in the main models presented in the manuscript and provided the detailed fit and loadings of the sensitivity analyses excluding the outlying classrooms in Appendix A9 and A10.

Exploratory factor analyses. The exploratory factor analyses were informative, in relation to the pattern of within- and cross-domain item correlations we saw in Table 1, as well as the alternative structures identified by prior studies. In 2009, the three-factor exploratory model had the best fit and met absolute fit for CFI (CFI= .96, NNFI= .91, RMSEA= .09; see Appendix A11). In 2014, a two-factor exploratory model was the best-fitting well-identified model, although not meeting absolute criteria (CFI= .94, NNFI= .90, RMSEA= .09; see again Appendix A11). Notably, in both samples, some exploratory factor models failed to converge or had negative residual variances. We discuss best fitting models in the next sections, considering both the traditional and bifactor structures for the same well-fitting number of factors because (as noted above) in exploratory factor analyses these traditional and bifactor solutions reflect

different rotations of equivalent-fitting solutions.

One-general/two-domain-specific bifactor (FACES 2009). In FACES 2009, where a three-factor solution was best fitting, its bifactor rotation mimicked the structure identified by Hamre and colleagues (2014). (See Appendix A12 for the loadings). That is, every item loaded ≥ .40 on the general factor. One of the two domain-specific factors mirrored Hamre et al.'s Cognitive Facilitation domain-specific factor, with the final three items (concept development, quality of feedback, and language modeling) all having loadings ≥ .40. Also consistent with Hamre et al.'s Proactive Management and Routines domain-specific factor, the highest-loading items on the other exploratory domain-specific factor were positive climate, negative climate, behavior management, and productivity, with loadings ranging from .31 to .55. The remaining three items had smaller loadings on both domain-specific factors (-.27 to .12), although they all loaded ≥ .40 on the general factor, again consistent with Hamre et al.'s results where these three items had zero or negative loadings on the domain-specific factor.

Traditional three-factor (FACES 2009). In the traditional three-factor rotation for 2009, a simple structure was evident, meaning that each item loaded ≥ .40 on one factor and < .40 on the other two factors (see again Appendix A12). The composition of the three factors differed, however, from the standard CLASS domains, but paralleled the pattern of loadings in the Hamre et al. bifactor rotation. The Instructional Support domain was evident, with the same three items as in the standard structure (concept development, quality of feedback, language modeling). However, the remaining seven items were grouped differently. One of the new factors—which we referred to as *Climate & Management*—had the positive climate, negative climate, behavioral management and productivity items with loadings ≥ .40. The other new factor—which we called *Sensitivity & Regard*—had the teacher sensitivity, regard for student perspectives, and instructional learning formats items with loadings ≥ .40.

One-general/one-domain-specific bifactor (FACES 2014). In 2014, the bifactor solution with one general and one domain-specific factor revealed all items but negative climate loading ≥ .40 on the general factor, but only the Instructional Support items loading on the specific factor (see Appendix A13).

Traditional two-factor (FACES 2014). The traditional two-factor solution in FACES 2014 generally combined the first two standard CLASS domains (see again Appendix A13), consistent with the structure identified by Cadima and colleagues (2016). That is, all of the first seven items except negative climate loaded ≥ .40 on one of the factors; negative climate loaded at .36. The Instructional Support items had small loadings on this factor (.04 and below). On the second factor, the three Instructional Support items had loadings ≥ .40, and the other items had loadings of .12 and below.

We next considered each of these four models, and the other models described above, in the confirmatory factor analyses.

Research Question 2: What Is the Best-fitting Dimensional Structure When Comparing Confirmatory Models?

Fit of traditional models. Fit indices for confirmatory factor models are shown in Table 2. Among the traditional models (top panel of Table 2), the two- and three-domain structures (Models 1a to 1c) had appreciably better fit than the one-factor traditional structure (Model 1d) in both FACES 2009 and FACES 2014. Across the two- and three-domain structures, however, just one index met absolute fit criteria and only for one model. That is RMSEA was at .08 for Model 1b, the alternative three-domain structure, in FACES 2009. This Model 1b from FACES 2009 was also the only structure that had a CFI value more than .01 larger than other traditional two- and three-factor structures. Based on this set of results, we featured below factor loadings from three traditional models, Models 1a, 1b, and 1c. Model 1a had the three standard CLASS

domains. Model 1b was the alternative structure better fitting in FACES 2009 (three-factor traditional structure, placing the first seven items on different factors than the standard scoring). Model 1c was the alternative equally well-fitting structure in FACES 2014 (two-factor traditional structure, combining together the Emotional Support and Classroom Organization items).

Fit of bifactor models. Turning to the bifactor models (bottom panel of Table 2), several fit indices met absolute criteria, however the only model that met criteria across all three indices (Model 2a in FACES 2014) had a negative residual variance consistent with the estimation problems reported by Hamre et al. The bifactor model reflecting the standard CLASS domain structure (Model 2a) failed to converge in FACES 2009 and had a negative residual variance in FACES 2014. The two best-fitting models without problems with convergence or identification were: (a) the Hamre et al. two-domain-specific bifactor structure combining together the first seven items (Model 2c); and, (b) the alternative three-specific-factor structure reflecting our exploratory results (Model 2b). In both samples, these models had CFI values that met absolute fit criteria as did RMSEA values in 3 of 4 cases. Although Model 2c had slightly higher fit in FACES 2014 and Model 2d slightly higher fit in FACES 2009, in neither case did the difference in CFI exceed .01 (exact ΔCFI available in the notes to Table 2). The final bifactor models (2d and 2e) showed appreciably worse fit. Because no bifactor model fit appreciably better than the Hamre et al. 2014 specification, we featured the Model 2c loadings below. Doing so facilitated comparison with the Hamre et al. 2014 published results.

Factor loadings. Table 3 provides the factor loadings for the better fitting models just discussed. In most cases, the loadings were ≥ .40 in each featured solution.

One exception was for the domain-specific Proactive Management and Routines factor of the bifactor model in FACES 2009, where no loadings were ≥ .40 and four were negative (see "PMR" labelled column in Table 3). These results were consistent with those found by Hamre

and colleagues in that they found five of the eight items loaded < .40 on the Proactive Management and Routines factor, two with negative loadings and one with a zero loading. Of note, the valences of loadings of our results were consistent with those of Hamre and colleagues. These valences also mirrored the alternative three-factor traditional structure that we identified. That is positive climate, negative climate, behavioral management and productivity loaded in the same direction on Proactive Management and Routines factor, whereas teacher sensitivity, regard for student perspectives, and instructional learning formats loaded in the opposite direction.

In the bifactor structure for FACES 2014, it was also the case that not all items loaded ≥ .40 on the general factor. The exceptions were the Instructional Support items, as well as negative climate. The factor loadings for the first domain-specific factor differed from FACES 2009 and Hamre et al., with just three of the Emotional Support and Classroom Organization items loading ≥ .40 (positive climate, teacher sensitivity, and behavior management).

The other exception to loadings being ≥ .40 involved negative climate in FACES 2014, which did not load at this level in any of the structures.

Research Question 3: How Intercorrelated Are the Dimensions?

In the traditional confirmatory factor analytic structures, the inter-factor correlations were generally moderate to high, although especially so for Emotional Support and Classroom Organization (results not tabled). For the raw scores of the three-domain structures, the correlations were $r_{12} = .76, r_{13} = .51, r_{23} = .51$ for the standard CLASS domains and $r_{12} = .69, r_{13} = .51, r_{23} = .48$ for the alternative structure in FACES 2009. In FACES 2014, they were $r_{12} = .79, r_{13} = .41, r_{23} = .47$ for the standard CLASS domains and $r_{12} = .86, r_{13} = .42, r_{23} = .48$ for the alternative structure. For the two-domain structure, the correlation was .54 in FACES 2009 and .46 in FACES 2014.

For the latent correlations estimated by Mplus, the values were $r_{12} = .94$, $r_{13} = .52$, $r_{23} = .57$ for the three standard CLASS domains and $r_{12} = .84$, $r_{13} = .57$, $r_{23} = .48$ for the alternative structure in FACES 2009. In FACES 2014, the values were $r_{12} = .94$, $r_{13} = .28$, $r_{23} = .29$ for the three standard CLASS domains and $r_{12} = .98$, $r_{13} = .26$, $r_{23} = .32$ for the alternative structure. For the two-domain structure, the latent correlation was .55 in FACES 2009 and .29 in FACES 2014.

Altogether, these results are consistent with prior studies finding relatively high inter-domain correlations, especially between Emotional Support and Classroom Organization where correlations ranged from .69 to .98 reflecting 48% to 96% shared variation (based on the square of the correlations). In contrast, correlations of these two domains with Instructional Support were lower, at .26 to .57, reflecting 7% to 32% shared variation.

Recall that the confirmatory bifactor model is specified with orthogonal factors, and thus its inter-factor correlations were constrained to zero.

Research Question 4: How Do Traditional and Factor Scorings Relate to Children's Academic and Socio-Emotional Outcomes?

We now turn to the regression models predicting children's spring assessment scores, controlling for the corresponding fall assessment score and other covariates. The results are summarized in Tables 4-6. A main takeaway from these results is that the associations of CLASS PreK scores with child outcomes were uniformly small and primarily non-significant in these national samples of children participating in Head Start. That is, only 12 of 240 coefficients (5%) had a *p*-value less than .05, and which associations were significant differed between the two samples. Regarding substantive size, the significant coefficients ranged in magnitude from .05 to .14, and some had signs of opposite valence than would be conceptually expected.

English language academic assessments. Beginning with the English language academic

assessments in Table 4, 5 of 80 coefficients (6%) were significant at $p < .05$, each with standardized valence of .10 or smaller in magnitude. In FACES 2009, the total raw score and Instructional Support raw score significantly positively associated with children's growth in Woodcock Johnson Letter Word scores. In FACES 2014, the Instructional Support raw score and the Cognitive Facilitation domain-specific factor significantly positively associated with gains in the Woodcock Johnson Applied Problems scores, but the Proactive Management and Routines domain-specific score significantly associated with *losses* in the Woodcock Johnson Letter Word scores. Altogether, 94% of coefficients (75 of 80) were nonsignificant, 96% (77 of 80) were below .10 in size, none exceeded .20 in size, and the average of the standardized coefficients was .01 in both samples.

Spanish language academic assessments. For the Spanish language academic assessments, just 1 of 80 associations was significant (1%), and it was negatively valanced: The Cognitive Facilitation domain-specific factor was associated with losses in the TVIP receptive vocabulary scores in FACES 2009 ($\beta = -.11$; Table 5). Altogether, 99% of coefficients (79/80) were nonsignificant, 71% (57 of 80) coefficients were smaller than .10 in size, all were below .20 in size, and the average of the coefficients was .00 in 2009 and -.02 in 2014.

Social-emotional assessments. For social-emotional assessments, 6 of 80 associations were significant (8%), but with different outcomes between the two samples (Pencil Tapping and teacher-reported social skills in FACES 2009; Leiter and teacher-reported behavior problems in FACES 2014). These significant associations ranged in magnitude from .07 to .14. One was valanced in a conceptually unexpected direction: The Proactive Management and Routines domain-specific score was associated with losses in Pencil Tapping impulse control scores. Altogether, 93% of coefficients (74/80) were nonsignificant, 85% (68 of 80) were .10 or smaller in size, none exceeded .20 in size, and the average of the coefficients was .02 in 2009 and .01 in

2014 (after reversing the signs for behavior problems).

In Appendix A14 we also showed results for regression models predicting gains in parent-reported social skills and behavior problems, which were available only in FACES 2009. Just 1 of the 20 coefficients was significant, but it was negative in sign: Greater Instructional Support was associated with losses in parent-reported child social skills. All coefficients were smaller than .10 in magnitude, averaging .00 (after reversing the signs for behavior problems).

Summary. In sum, just 5% of all coefficients (12 of 240) was significant. The potential that these reflected chance, given a 5% alpha level, was reinforced by the fact that which associations were significant did not replicate between the two FACES samples and some associations were of opposite sign than would be expected conceptually. The significant coefficients were also small in size, with standardized coefficients ranging in magnitude from .05 to .14. In other words, Head Start children attending classrooms in the spring that scored a full standard deviation higher on a CLASS domain generally averaged a non-significant 5 to 14 hundredths of a standard deviation higher on their spring assessments, adjusting for their starting level on the assessments in the fall and their demographic characteristics.

Discussion

The current study replicated and extended recent factor analyses of the CLASS PreK, focusing specifically on the measure's domain structure within two nationally representative Head Start samples (FACES 2009 and 2014). Like earlier studies, we found that whereas the standard CLASS three-domain structure fit better than a one-domain structure, it did not meet absolute fit criteria and other structures were equally or better fitting.

Specifically, we found that a revised three-domain structure fit better than the standard CLASS structure in FACES 2009, and both three-domain structures as well as a two-domain structure fit equally well in FACES 2014. Consistent with prior studies, these alternative

structures adjusted the placement of the Emotional Support and Classroom Organization items. The revised three-domain structure reorganized these items into a Climate & Management domain (positive climate, negative climate, behavioral management, productivity) and a Sensitivity & Regard domain (teacher sensitivity, regard for student perspectives, and instructional learning formats), while retaining the Instructional Support domain. The two-domain structure combined the Emotional Support and Classroom Organization items and retained the Instructional Support domain. A bifactor structure identified by Hamre and colleagues (2014) with two domain-specific factors, the first combining the Emotional Support and Classroom Organization items and the second reflecting Instructional Support items, had adequate fit on some indices, although like Hamre et al. we found that not all items loaded meaningfully on a domain-specific factor and bifactor structures had problems with convergence and identification.

Even with these new structures, our results reinforced weaknesses in the evidence base that Hamre and colleagues (2014) had summarized, including moderate to high factor intercorrelations, especially for the Emotional Support and Classroom Organization domains which shared about 50% or more variation. Scores based on all solutions also had small and typically nonsignificant associations with children's gains in academic and social-emotional skills during the Head Start year.

Placing Results in Context of Prior Studies

Our results replicate and extend prior studies. One consistent theme, echoing prior studies, is that the Instructional Support factor is most consistently revealed. The Emotional Support and Classroom Organization domains are less distinct, with equally or better fitting structures that combine the two domains together or show patterns of item loadings different from the standard scoring and with high correlations between the domain scores. At a big picture

level these conclusions replicate between FACES 2009 and 2014, although some specific results differ. The alternative structure we identified that was better fitting in FACES 2009 mirrored the loading pattern found by Hamre et al. 2014 whose data were collected in 2008 and 2009, a similar time period as FACES 2009. In FACES 2014, however, this structure was equally well fitting as other structures, including a two-domain structure combining together all of the Emotional Support and Classroom Organization items that has been evident in other prior studies (e.g., Cadima et al., 2016). The negative climate item also showed consistently lower loadings in FACES 2014 than FACES 2009, consistent with prior studies that excluded this item (e.g., Pakarinen et al., 2010).

These varying results may reflect any number of differences between study designs and samples, including those corresponding to historical and programmatic changes. One salient trend between 2009 and 2014, for instance, was efforts to increase teacher education levels in Head Start. We see this trend reflected in an increase from about half to nearly three-quarters of teachers having a bachelor's degree in 2009 versus 2014. Implementation of the Head Start DRS also expanded between 2009 and 2014, which would have exposed more teachers to CLASS PreK and would have altered the underlying population of programs as some had to recompete. We encourage future studies to probe such possible reasons for differences in factor structures, including with psychometric meta-analyses and tests for measurement invariance (Fujimoto, Gordon, Peng & Hofer, 2018; Millsap, 2011). We also encourage continuous measure development that is flexible to potential variation in the definition and expression of key constructs across subcontexts and subpopulations including by space and time (AERA/APA/NCME, 2014; Davidson et al., 2018; Gordon, 2015).

Our findings of small associations with children's gains is also consistent with prior studies of CLASS PreK and other classroom quality measures. For instance, in an early meta-

analysis of twenty published studies covering a range of quality measures and child outcomes, Burchinal and colleagues (2011) estimated an overall effect size of .11. In a more recent re-analysis of secondary datasets that looked for possible threshold or dosage effects, the authors found little evidence of non-linear associations with the largest effect sizes being .08 (Burchinal et al., 2016). In a more recent comprehensive meta-analysis focused on CLASS PreK, Perlman and colleagues (2016) similarly found effect sizes of at most .09 in their meta-analysis of linear associations. Our results replicate and extend this evidence by being focused on recent cohorts of Head Start children, thereby offering results specifically relevant to the Head Start DRS, and by looking at scores produced for a range of alternative factor structures. We also included results for a broader array of outcomes and subsamples than have some prior studies, including a direct assessment measure of executive functioning and Spanish speaking subpopulations.

Interpretation of Results

When interpreting these results, it is important to recognize that the lack of a clear domain structure for the CLASS PreK items does not necessarily mean that the underlying theoretical frameworks on which the CLASS was based are incorrect. Rather, it may be that different operationalizations of these frameworks are needed.

Our results solidify evidence related to how well the CLASS PreK operationalizes not only its original three domains but also its more recently proposed general/specific bifactor structure. As Hamre (2014) noted in a review of the theoretical basis of the CLASS, the bifactor model's ability to separate a domain-general responsive teaching factor from domain-specific factors aligns with decades of developmental research. Teachers' general responsiveness may reflect their cross-cutting ability to stay in tune with children's cues, individualize support, and scaffold feedback. Specific aspects of teaching might include their particular management styles and curricular strategies. However, Hamre (2014, p. 225) did not feature the bifactor model in

her review, instead focusing on the three standard CLASS domains, because the PreK bifactor structure had not been replicated (Hamre, 2014, p. 225). Our study helps to build such needed replication, adding to recent examinations of the PreK bifactor structure in other countries and of similar structures in versions of the CLASS for elementary and secondary classrooms. This emerging body of evidence does not point to a single best alternative structure. However, it does demonstrate that, as currently operationalized, the CLASS PreK items do not well reflect the three standard domains of Emotional Support, Classroom Organization, and Instructional Support nor the alternative general/specific bifactor structure.

Although numerous limitations in the measure's operationalization might be addressed, including low inter-rater reliability and considerable cross-cycle score fluctuations (Burchinal, 2018; Mashburn, 2017), we focus our attention on the central topic of this paper, number of domains. Sharpening the definition of each domain would be an important starting point to establishing empirical support, including for the original three domains as well as the emerging general/specific domains and alternative traditional structures. After establishing such definitions, explicitly making a case for the extent to which each item reflects them would be an important next step. Aiming for items that well reflect one and only one domain would best support simple factor structures and reduce inter-factor correlations. Such activities can be accomplished through expert ratings and panels to elicit views from multiple vantage points (AERA/APA/NCME, 2014; Davidson et al., 2018; Wolfe & Smith, 2007).

This foundational definition-refinement and item-mapping process may be particularly helpful given the structure of the CLASS manual which relies on lengthy narratives describing hallmark indicators and markers of low, middle, and high scores for each item. Extending the expert review process to more explicitly parse apart and rate to which domains the statements of these narratives apply could improve operationalization of the underlying concepts. Where

warranted, the considerable content in each narrative might be turned into new items whose scores are retained and analyzed, covering each refined domain definition. The CLASS manual's indicators and markers might be a starting point, since these are used to guide ratings but are not numerically recorded nor incorporated mathematically in standard scoring. Rather, currently, raters make individualized mental assessments of how to weigh these indicators and markers. As the manual notes "...the CLASS is *not a checklist* and observers should view the dimensions as holistic descriptions of classrooms that fall in the low, middle, or high range. In many cases, it is not necessary to see indicators of all markers presented in the description of a given range to assign a score in that range" (Pianta et al., 2008, p. 15). Some scholars might expect the resulting more granular coding to be more time consuming and taxing than more inferential and global scoring and that human inference may be imperfect but still best for summarizing the complex activities and interactions that take place in early childhood classrooms. To inform these possibilities, modern data collection strategies and statistical modeling could be used to develop and compare scoring alternatives (numerically capturing and mathematically summarizing indicators and markers versus allowing raters to mentally assimilate the indicators and markers in individualized ways) while documenting their relative burden and accuracy. While doing so, future research might tap into rich literatures in other subfields on the tension between reliability and validity in rater-mediated measures, including those based on more and less inferential scoring (e.g., Elliot, 2005; Wind & Peterson, 2018).

Such a revision process might also address the restricted range of CLASS items, which may contribute to estimation problems evident in our and prior studies. Attending to writing items that better cover the easier end of instruction-related continua and the harder end of emotion- and organization-related continua, might not only better support psychometric models of the theoretical domains but also improve variation for capturing teachers' growth in practices

and predicting their students' development. In other words, measure refinement could consider whether the narrative content of the high (6, 7) categories capture the maximum conceptual aspects of emotional support and classroom organization, or whether additional content might be added to allow more room for growth. Likewise, at the same time that high scores on instructional support might be ideal aspirational goals for classrooms, exploring narrative extensions that could better distinguish among the teachers currently clustered in the low (1,2) scores could improve variation for predicting child growth and allow for capturing teacher improvement as they build skills when starting from the lower end of instruction.

Will the bifactor model be useful in such efforts to improve operationalization of theoretically-informed measures of classroom quality? As other scholars have noted, the bifactor model is not a panacea (Hafen et al., 2015; Leyva et al., 2015). One limitation is the restriction of factors to be uncorrelated, which is beneficial for separating unique aspects of classroom quality (thereby circumventing the problem of moderate to high empirical cross-domain correlations) but may not reflect conceptually or practically interpretable domains (because in reality teacher behaviors and practices are interrelated). The bifactor structure is also unfamiliar to many practitioners and scholars. As Hafen and colleagues (2015) note, as much as the idea of general and specific domains are theoretically appealing, additional effort is needed to explain the bifactor model to non-technical audiences and to produce factor scores for practice and policy uses. Rather than relying on statistical models to parse apart general and specific aspects of the small number of CLASS items, the approaches described above to better define and differentiate domains and their operationalization might help produce items with empirical support for the several conceptually expected domains of early childhood classroom quality in traditional (rather than bifactor) structures.

Implications

Our results have important implications for research, practice, and policy. In relation to research, the process just described of measure refinement can inform the conceptual basis of the CLASS PreK, as greater precision in defining domains and their operationalization can feed back into theory building. Improved operationalization can also reduce measurement error and increase variation, both of which can support modeling how classroom quality associates with children's development and potentially identify stronger associations with certain outcomes. These improved operationalizations and strengthened evidence base can also support practice and policy by providing teachers and leaders with even more precise guidance and for better documentation of growth in classroom quality. When considering the current evidence, decision makers should follow modern measurement guidelines to weigh the full body of evidence in relation to each possible use (AERA/APA/NCME, 2014). The evidence offered here suggests that the CLASS PreK should not be described to researchers, practitioners, and policymakers as though the way that it operationalizes the three standard CLASS domains is well validated empirically. We would also conclude that our evidence does not support using the three standard domains in high stakes ways, such as in the Head Start DRS.

Limitations

Our study was limited. Although using two national samples of Head Start classrooms offered ideal evidence for Head Start decision makers, whether and how the results would generalize to classrooms and programs not funded by Head Start is uncertain. We also focused on the first Head Start year, with intervals averaging 6 months between fall and spring assessments. Although this time span aligns with policy decision-making about how classroom quality associates with gains during typical exposure to Head Start, results might differ in the second Head Start year or for longer term outcomes. Our sample sizes also were large overall, although the sample sizes of children assessed in Spanish were smaller leading to imprecise

estimates for this subpopulation. Moreover, although we examined numerous measures of children's development, only one direct assessment was available for the social-emotional domains. Including additional direct assessment measures (Lipsey et al., 2017) would improve future studies, as would ongoing efforts to improve measures of young children's development across physical, academic, and social-emotional domains. It is also the case that, although FACES followed CLASS PreK certification procedures (and exceeded them with live reliability checks before and during data collection), the substantial rater error left by within-one percentage agreement is increasingly recognized as problematic for many uses of the scores (Burchinal, 2018). CLASS PreK scores also were gathered on a single day in the spring, at the end of the Head Start year, again reflective of common practice but not ideal for capturing likely fluctuations over time.

Conclusions

In short, our results encourage continued efforts to improve existing and to develop new measures of early childhood classroom quality. As already noted, our results do not necessarily mean that other operationalizations of the theoretical framework underlying the CLASS PreK would not better identify the three expected domains and their anticipated associations with child outcomes. The CLASS PreK may also have value in certain uses, such as when offering formative verbal feedback to teachers about areas for improvement. What the results do mean is that as currently operationalized, the CLASS PreK does not well differentiate its three expected domains, and, that the resulting scores do not demonstrate the expected associations across children's outcomes. These results have important implications for high stakes use, such as those in the Head Start DRS which was predicated on an assessment of classroom quality that was "linked to positive child development and later achievement" (Public Law 110-134).

Broadly, our results are relevant to the way that CLASS PreK domains have been widely

used in research and written into high stakes policy decisions including the Head Start DRS. Our replication and extension of alternative factor structures for the CLASS PreK caution against such high stakes uses of the standard domains. We encourage future policy decision-making about classroom quality measures to follow modern measurement standards including a careful assessment of the full body of evidence specifically related to a particular use and the building in of flexibility for ongoing accumulation of evidence and continuous measure improvement (AERA/APA/NCME, 2014; Davidson et al., 2018; Gordon, 2015; Mashburn, 2017).

References

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

Bihler, L., Agache, A., Kohl, K., Willard, J. & Leyendecker, B. (2018). Factor analysis of the Classroom Assessment Scoring System replicates the three domain structure and reveals no support for the bifactor model in German preschools. *Frontiers in Psychology, 9*, 1232.

Bishara, A. J. & Hittner, J. B. (2017). Confidence intervals for correlations when data are not normal. *Behavioral Research Methods, 49*, 294-309.

Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist, 57,* 111–127.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research: Second edition.* New York, NY: Guilford.

Burchinal, M. (2018). Measuring early care and education quality. *Child Development Perspectives, 12*, 3-9.

Burchinal, M., Kainz, K., & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle. (Eds.), *Quality measurement in early childhood settings* (pp. 11–31). Baltimore, MD: Brookes Publishing.

Burchinal, M., Zaslow, M., & Tarullo, L. (2016). Quality thresholds, features, and dosage in early care and education: Secondary data analysis of child outcomes. *Monographs of the Society for Research in Child Development*.

Cadima, J., Verschueren, K., Leal, T. & Guedes, C. (2016). Classroom interactions, dyadic teacher-child relationships, and child self-regulation in socially disadvantaged young children. *Journal of Abnormal Child Psychology, 44*, 7-17.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science.* Oakland, CA: University of California Press.

Colwell, N., Gordon, R. A., Fujimoto, K., Kaestner, R., & Korenman, S. (2013). New evidence on the validity of the Arnett Caregiver Interaction Scale: Results from the Early Childhood Longitudinal Study-Birth Cohort. *Early Childhood Research Quarterly, 28,* 218–233.

Davidson, L., Crowder, M., Gordon, R. A., Domitrovich, C., Brown, R., & Hayes, B. (2018). A continuous improvement approach to social and emotional competency measurement. *Journal of Applied Developmental Psychology*, *55*, 93-106.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to 'Do as I say, not as I do.' *Developmental Psychobiology*, 29, 315–334.

Duncan, S. E., & DeAvila, E. (1998). *Preschool Language Assessment Survey 2000.* Monterey, CA: CTB/McGraw-Hill.

Dunn, L. M., Padilla, E. R., Lugo, D.E. & Dunn, L. M. (1986). *Test de Vocabulario en Imagenes Peabody.* Circle Pines, MN: American Guidance Service.

Dunn, L. M., Dunn, L. L., & Dunn, D. M. (2006). *Peabody Picture Vocabulary Test, Fourth Edition*. Circle Pines, MN: American Guidance Service.

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in g-factor models: Explanations and alternatives. *Psychological Methods, 22*, 541-562.

Elliot, N. (2005). *On a Scale: A Social History of Writing Assessment in America*. New York, NY: Peter Lang.

Elliott, S. N., Gresham, F. M., Freeman, T., & McCloskey, G. (1988). Teacher and observer ratings of children's social skills: Validation of the social skills rating system. *Journal of Psychoeducational Assessment, 6,* 152–161.

Entwisle, D. R., Alexander, K. L., Cadigan, D. & Pallis, P. M. (1987). The emergent academic self-image of first graders: Its response to social structure. *Child Development, 58,* 1190-1206.

Fujimoto, K. A., Gordon, R. A., Peng, F., & Hofer, K. G. (2018). Category functioning of the ECERS-R across eight data sets. *AERA Open, 4*, 1-16.

Gibbons, R. D., Bock, D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007) Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31(4),* 4-19.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika, 57,* 423–436.

Gordon, R. A. (2015). Measuring constructs in family science: How can item response theory improve precision and validity. *Journal of Marriage and Family, 77*, 147-176.

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the Classroom Assessment Scoring System – Secondary. *Journal of Early Adolescence, 35*, 651-680.

Hamre, B. K. (2014). Teachers' daily interactions with children: An essential ingredient in effective early childhood programs. *Child Development Perspectives, 8*, 223-230.

Hamre, B., Hatfield, B., Pianta, R., & 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.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis.* Boca Raton, FL: Chapman & Hall/CRC Press.

Hindman, A. H., Pendergast, L. L., & Gooze, R. A. (2016). Using bifactor models to measure teacher-child interaction in early childhood: Evidence from the Caregiver Interaction Scale. *Early Childhood Research Quarterly, 36,* 366-378.

Hu, B. Y., Fan, X., Gu, C., & Yang, N. (2016). Applicability of the Classroom Assessment Scoring System in Chinese preschools based on psychometric evidence. *Early Education and Development, 27*, 714-734.

Kang, Y., McNeish, D. M., & Hancock, G. R. (2016). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement, 76*, 533-561.

Klein, A. K., Carlson, B. L., Aikens, N., Bloomenthal, A., West, J., Malone, L., Moiduddin, E., Hepburn, M., Skidmore, S., Bernstein, S., Kelly, A., Hurwitz, F., & Lim, G. (2017). *Head Start Family and Child Experiences Survey (FACES) 2014: User's Manual.* Washington, DC: Mathematica Policy Research.

Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D., & Zald, D. H. (2017). A hierarchical causal taxonomy of psychopathology across the life span. *Psychological Bulletin, 143*, 142-186.

Leyva, D., Weiland, C., Barata, M., Yoshikawa, H., Snow, C., Treviño, & Rolla, A. (2015). Teacher-child interactions in Chile and their associations with prekindergarten outcomes. *Child Development, 86*, 781-799.

Lipsey, M., Nesbitt, K. T., Farran, D. C., Dong, N., Fuhs, M. W., & Wilson, S. J. (2017).

Learning-related cognitive self-regulation measures for prekindergarten children: A comparative evaluation of the educational relevance of selected measures. *Journal of Educational Psychology, 109*, 1084-1102.

Longobardi, C., Pasta, T., Marengo, D., Prino, L. E. & Settanni, M. (2018). Measuring quality of classroom interactions in Italian primary school: Structural validity of the CLASS K-3. *Journal of Experimental Education*. (Online first).

Madill, R., Gest, S., & Rodkin, P. C. (2013). A bifactor model of the CLASS: Associations with children's sense of relatedness and teachers' approaches to managing behavior and learning. 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.

Malone, L., Carlson, B. L., Aikens, N., Moiduddin, E., Klein, A. K., West, J., Kelly, A., Meagher, C., Bloomenthal, A., Hulsey, L., & Rall, K. (2013). *Head Start Family and Children Experiences Survey, 2009: User's Manual*. Princeton, NJ: Mathematica Policy Research.

Martin, N. & R. Brownell. (2012). Receptive One-Word Picture Vocabulary Test–4th Edition: Spanish-Bilingual Edition. Novato, CA: Academic Therapy Publications.

Mashburn, A. J. (2017). Evaluating the validity of classroom observations in the head start designation renewal system. *Educational Psychologist, 52,* 38-49.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Muthén, L. K. & Muthén, B. O. (2012). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

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 & Development*, *21,* 95–124.

Perlman, M., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. S. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the Classroom Assessment Scoring System) in early childhood education and care settings and child outcomes. *PLOS One, 11*, 1-33.

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

QRIS Compendium. (2017). *QRIS Compendium.* Retrieved from http://qriscompendium.org/

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society (Series A), 169,* 805–827.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47(5),* 667-696.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society: Series B (statistical methodology), 71*, 447-466.

Roid, G.H., & Miller, L. J. (1997). *Leiter International Performance Scale Revised, Examiner Rating Scale (Leiter-R)*. Lutz, FL: Psychological Assessment Resources.

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 (CLASS K-3). *Early Education and Development, 25*, 894-914.

Sandilos, L. E., Shervey, S. W., DiPerna, J. C., Lei, P., & Cheng, W. (2017). Structural validity of CLASS K-3 in primary grades: Testing alternative models. *School Psychology Quarterly, 32*, 226-239.

Skrondal, A. & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66 (4)*, 563-575.

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly, 22,* 173–187.

Von Suchodoletz, A., Fäsche, A., Gunzenhauser, C., & Hamre, B. K. (2014). A typical morning in preschool: Observations of teacher-child interactions in German preschools. *Early Childhood Research Quarterly, 29*, 509-519.

Weiland, C., Ulvestad, K., Sachs, J. & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly, 28*, 199-209.

Wind, S. A. & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*, 161-192.

Wolfe, E. W. & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models. *Journal of Applied Measurement, 8*, 1-27.

Woodcock, R.W., Muñoz-Sandoval, A.F., McGrew, K., Mather, N., & Schrank, F. (2004). *Batería III Woodcock-Muñoz.* Itasca, IL: Riverside Publishing.

Woodcock, R.W., McGrew, K. & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside Publishing.

Zill, N. (1990). *Behavior Problems Index based on parent report.* Washington, DC: Child Trends.

Table 1
Descriptive Statistics and Intercorrelations of CLASS Items

		FACES 2009														
								Correlations								
Domain	Item[a]	M	SD	Min	Max	Skew	Kurt	1	2	3	4	5	6	7	8	9
Emotional Support	CLASS1: Positive Climate	5.34	0.68	2.33	7.00	-0.02	0.19	-								
	CLASS2: Negative Climate (R)	6.75	0.47	2.33	7.00	**-3.52**	**20.43**	.42	-							
	CLASS3: Teacher Sensitivity	4.64	0.65	2.67	6.33	-0.31	-0.16	**.62**	.40	-						
	CLASS4: Regard for Student Perspectives	4.47	0.67	2.00	6.25	-0.55	0.58	.47	.37	**.73**	-					
Classroom Organization	CLASS5: Behavior Management	5.02	0.73	2.50	6.75	-0.49	0.00	**.58**	.47	**.60**	.46	-				
	CLASS6: Productivity	4.93	0.80	2.00	7.00	-0.36	0.23	**.51**	.42	**.54**	**.50**	**.56**	-			
	CLASS7: Instructional Learning Formats	4.00	0.81	1.75	6.00	-0.13	-0.55	.45	.21	**.62**	**.55**	.48	.47	-		
Instructional Support	CLASS8: Concept Development	2.12	0.67	1.00	4.25	0.52	0.15	.41	.19	.32	.29	.34	.38	.31	-	
	CLASS9: Quality of Feedback	2.30	0.69	1.00	5.00	0.73	0.63	.44	.29	.38	.27	.36	.36	.34	**.76**	-
	CLASS10: Language Modeling	2.45	0.80	1.00	5.00	0.76	0.76	.38	.37	.49	.42	.37	.40	.46	**.61**	**.76**
		FACES 2014														
								Correlations								
Domain	Item[a]	M	SD	Min	Max	Skew	Kurt	1	2	3	4	5	6	7	8	9
Emotional Support	CLASS1: Positive Climate	5.48	0.72	3.00	7.00	-0.38	0.28	-								
	CLASS2: Negative Climate (R)	6.75	0.42	3.67	7.00	**-2.25**	**6.15**	.37	-							
	CLASS3: Teacher Sensitivity	4.95	0.79	2.00	7.00	-0.36	0.07	**.74**	.28	-						
	CLASS4: Regard for Student Perspectives	4.70	0.78	1.67	7.00	-0.39	0.05	**.59**	.19	**.70**	-					
Classroom Organization	CLASS5: Behavior Management	5.10	0.80	2.33	7.00	-0.39	-0.24	**.68**	.36	**.74**	**.59**	-				
	CLASS6: Productivity	4.93	0.84	2.25	7.00	-0.29	-0.40	**.63**	.18	**.64**	**.58**	**.65**	-			
	CLASS7: Instructional Learning Formats	4.25	0.90	1.50	6.33	-0.33	-0.25	**.51**	.15	**.54**	.46	.49	**.67**	-		
Instructional Support	CLASS8: Concept Development	2.26	0.93	1.00	6.33	**1.02**	1.45	.22	-.03	.26	.25	.22	.21	.31	-	
	CLASS9: Quality of Feedback	2.53	1.00	1.00	6.75	0.80	0.68	.19	.02	.24	.22	.20	.20	.27	**.74**	-
	CLASS10: Language Modeling	2.55	0.94	1.00	6.25	0.69	0.45	.19	-.02	.28	.26	.22	.23	.27	**.62**	**.82**

Note. n = 370 (2009) and 641 (2014) classrooms. Skew = skewness. Kurt = Kurtosis. Skewness with absolute values above 1 and Kurtosis with absolute values above 2 were bolded. Shaded correlations fall within the standard CLASS domains. Correlations at or above .50 are bolded. All values were weighted by the classroom-level sampling weight. [a] For readers familiar with CLASS PreK, we note that we use the term items for what the CLASS PreK calls dimensions. We do so because we treat the 10 scores as the items that load on factors, as have other factor analytic studies of the CLASS PreK. Doing so avoids potential confusion given in factor analytic studies the term dimension is often used to refer to the latent construct measured by a factor.

Table 2
Fit Indices for Factor Analytic Models

	FACES 2009				FACES 2014			
	Fit Indices			Convergence &	Fit Indices			Convergence &
Model	CFI	NNFI	RMSEA	Estimation Issues	CFI	NNFI	RMSEA	Estimation Issues
Traditional Factor Analysis								
1a 3-Domain (standard) [a]	.91	.88	.10		.94	.92	.09	
1b 3-Domain (alternative) [a]	.94	.92	**.08**		.94	.91	.09	
1c 2-Domain (combine ES and CO) [a]	.91	.88	.10		.94	.91	.09	
1d Single domain	.71	.62	.18		.64	.53	.20	
Bifactor Analysis								
2a 3-Specific Factors (standard)	-	-	-	Failed to converge	**.98**	**.96**	**.06**	Negative residual variance
2b 3-Specific Factors (alternative)[b]	**.96**	.92	**.08**		**.96**	.94	**.08**	
2c 2-Specific Factors (combine ES and CO)[b]	**.95**	.91	.09		**.97**	.94	**.07**	
2d 2-Specific Factors (ES and CO, adjusted)	.88	.81	.13		.93	.89	.10	Negative residual variance
2e 1-Specific Factor (IS only)	.93	.90	.10		.94	.92	.09	

Note. n = 370 (2009) and 641 (2014) classrooms. Models were weighted by the classroom-level sampling weight, and scores were reversed for the negatively oriented negative climate item. CFI = comparative fit index. NNFI = nonnormed fit index. RMSEA = root-mean-square error of approximation. Bolded values meet criteria of NNFI ≥ .95, CFI ≥ .95, RMSEA ≤ .08. Among 3-factor traditional and 3-domain-specific bifactor structures, the "standard" models have three (domain-specific) factors comprised of Items 1-4, 5-7, and 8-10; the "alternative" models have three (domain-specific) factors comprised of Items 1,2,5,6; 3,4,7; and, 8-10. The "combine ES and CO" models have two (domain-specific) factors comprised of Items 1-7, and 8-10. The "one-factor traditional" model has a single factor with all Items 1-10. The "ES and CO, adjusted" model has two domain-specific factors comprised of Items 1,2,5,6 and 8-10. The "1-Specific Factor (IS only)" model has one domain-specific factor with items from the Instructional Support domain (Items 8-10). [a] The differences in CFI among Models 1a, 1b, and 1c were below .01 in FACES 2014 (from .001 to .005); for FACES 2009, Model 1b exceeded Model 1a by .03 and Model 1c by .034. [b] The difference in CFI between Model 2b and Model 2c is .007 in FACES 2009 and .004 in FACES 2014.

Table 3

Confirmatory Factor Loadings from Hamre et al. Two-Domain-Specific Bifactor Structure and Traditional Three-Factor Standard and Alternative Structures

	FACES 2009										
	Bifactor (Orthogonal)			Traditional CFA (Oblique)							
	General	Domain-Specific		Standard 3-Factor			Alternative 3-Factor			Combine ES & CO 2-Factor	
Item	RT	PMR	CF	ES	CO	IS	CM	SR	IS	ESCO	IS
CLASS1: Positive Climate	**.73**	*-.14*		**.71**			**.76**			**.72**	
CLASS2: Negative Climate (R)	**.53**	*-.24*		**.50**			**.56**			**.51**	
CLASS3: Teacher Sensitivity	**.87**	.23		**.89**				**.92**		**.87**	
CLASS4: Regard for Student Perspectives	**.75**	.32		**.77**				**.79**		**.75**	
CLASS5: Behavior Management	**.75**	*-.27*			**.74**		**.77**			**.72**	
CLASS6: Productivity	**.69**	*-.15*			**.71**		**.72**			**.69**	
CLASS7: Instructional Learning Formats	**.67**	.17			**.69**			**.68**		**.68**	
CLASS8: Concept Development	**.44**		**.64**			**.80**			**.80**		**.80**
CLASS9: Quality of Feedback	**.47**		**.87**			**.94**			**.94**		**.94**
CLASS10: Language Modeling	**.56**		**.56**			**.81**			**.80**		**.81**

	FACES 2014										
	Bifactor (Orthogonal)			Traditional CFA (Oblique)							
	General	Domain-Specific		Standard 3-Factor			Alternative 3-Factor			Combine ES & CO 2-Factor	
	RT	PMR	CF	ES	CO	IS	CM	SR	IS	ESCO	IS
CLASS1: Positive Climate	**.70**	**.45**		**.83**			**.83**			**.83**	
CLASS2: Negative Climate (R)	.20	.33		.33			.34			.33	
CLASS3: Teacher Sensitivity	**.74**	**.50**		**.90**				**.89**		**.89**	
CLASS4: Regard for Student Perspectives	**.64**	.38		**.76**				**.76**		**.75**	
CLASS5: Behavior Management	**.69**	**.46**			**.84**		**.83**			**.83**	
CLASS6: Productivity	**.82**	.11			**.80**		**.77**			**.77**	
CLASS7: Instructional Learning Formats	**.84**	*-.17*			**.67**			**.64**		**.64**	
CLASS8: Concept Development	.34		**.68**			**.76**			**.76**		**.76**
CLASS9: Quality of Feedback	.30		**.95**			**.98**			**.97**		**.98**
CLASS10: Language Modeling	.32		**.76**			**.84**			**.84**		**.84**

Note. n = 370 (2009) and 641 (2014) classrooms. Values are standardized factor loadings from models that applied the classroom-level sampling weights. (R) reflects our reversal of scores for the negatively oriented negative climate. Bolded values are > |.40|. Italicized values are negative. RT = Responsive Teaching. PMR = Proactive Management and Routines. CF = Cognitive Facilitation. ES = Emotional Support. CO = Classroom Organization. IS = Instructional Support. CM = Climate & Management. SR = Sensitivity & Regard. ESCO = Emotional Support & Classroom Organization.

Table 4
Standardized Regression Coefficients for Child Academic Outcomes, English Version

FACES 2009				
	PPVT	Woodcock Johnson		
CLASS Scores		Applied Problems	Letter Word	Spelling
Total Raw Score	0.01 (0.02)	-0.02 (0.03)	**0.06 (0.02)***	0.00 (0.03)
Raw Average (3d Standard)				
Emotional Support (Items 1,2,3,4)	-0.04 (0.04)	-0.02 (0.04)	-0.02 (0.04)	-0.06 (0.04)
Classroom Organization (Items 5,6,7)	0.03 (0.04)	0.02 (0.04)	0.03 (0.04)	0.06 (0.05)
Instructional Support (Items 8,9,10)	0.04 (0.03)	-0.02 (0.04)	**0.05 (0.03)***	0.00 (0.04)
Raw Average (3d Alternative) [a]				
Climate & Management (Items 1,2,5,6)	0.01 (0.03)	-0.03 (0.03)	0.04 (0.04)	-0.02 (0.04)
Sensitivity & Regard (Items 3,4,7)	-0.03 (0.03)	0.03 (0.04)	-0.03 (0.04)	0.02 (0.05)
Raw Average (2d Alternative) [a]				
Combine ES & CO (Items 1-7)	-0.02 (0.03)	0.00 (0.03)	0.01 (0.03)	0.00 (0.04)
Confirmatory Bifactor				
General	0.00 (0.02)	-0.01 (0.03)	0.04 (0.03)	0.00 (0.03)
Proactive Mgt. & Routines (Items 1-7)	-0.02 (0.02)	0.01 (0.03)	-0.03 (0.03)	-0.01 (0.03)
Cognitive Facilitation (Items 8-10)	0.03 (0.02)	-0.02 (0.03)	0.03 (0.02)	0.00 (0.03)
FACES 2014				
	PPVT	Woodcock Johnson		
		Applied Problems	Letter Word	Spelling
Total Raw Score	-0.01 (0.05)	0.05 (0.03)	0.03 (0.05)	0.06 (0.04)
Raw Average (3d Standard)				
Emotional Support (Items 1,2,3,4)	0.00 (0.08)	-0.02 (0.06)	-0.08 (0.08)	-0.08 (0.07)
Classroom Organization (Items 5,6,7)	-0.04 (0.10)	-0.01 (0.06)	0.07 (0.08)	0.12 (0.07)
Instructional Support (Items 8,9,10)	0.03 (0.04)	**0.10 (0.03)***	0.04 (0.05)	0.02 (0.05)
Raw Average (3d Alternative) [a]				
Climate & Management (Items 1,2,5,6)	-0.09 (0.07)	-0.02 (0.06)	-0.13 (0.09)	0.05 (0.08)
Sensitivity & Regard (Items 3,4,7)	0.06 (0.07)	-0.02 (0.06)	0.13 (0.09)	-0.01 (0.09)
Raw Average (2d Alternative) [a]				
Combine ES & CO (Items 1-7)	-0.03 (0.05)	-0.04 (0.04)	-0.01 (0.06)	0.04 (0.05)
Confirmatory Bifactor				
General	-0.01 (0.05)	-0.01 (0.03)	0.05 (0.05)	0.05 (0.04)
Proactive Mgt. & Routines (Items 1-7)	-0.02 (0.04)	0.00 (0.03)	**-0.09 (0.04)***	-0.01 (0.04)
Cognitive Facilitation (Items 8-10)	0.02 (0.04)	**0.10 (0.03)***	0.04 (0.05)	0.04 (0.05)

Note. Values are standardized coefficients (and standard errors) from five mixed models regressing children's spring scores on their classroom's spring CLASS score(s). The five models reflect the total score and four sets of scores. Covariates included the children's fall scores on the respective outcome as well as their gender and race-ethnicity, low-income, and disability status, whether their mother had a high school degree or less, their fall age cohort (3- or 4-year old), their age in months at the time of the spring assessment, and the months elapsed between the fall and spring assessments. The models included random intercepts at the classroom level and applied the child weights at the child-level and the classroom weights at the classroom level using the size approach to scaling the multi-level weights. Item-level missing values were imputed using multiple imputation (total *n* children = 1,984 and *n* classrooms = 361 in 2009; *n* children = 833 and *n* classrooms = 189 in 2014). * $p < .05$ (also bolded). [a] Results for the average raw score of Items 8-10 provided once, in the row labelled Instructional Support, to avoid duplication. ES = Emotional Support. CO = Classroom Organization. Mgt. = Management.

Table 5
Standardized Regression Coefficients for Child Academic Outcomes, Spanish Version

CLASS Scores	TVIP	Woodcock-Muñoz		
		Applied Problems	Letter Word	Spelling
FACES 2009				
Total Raw Score	0.02 (0.06)	-0.02 (0.07)	-0.01 (0.07)	-0.10 (0.06)
Raw Average (3d Standard)				
Emotional Support (Items 1,2,3,4)	-0.01 (0.06)	-0.05 (0.10)	-0.15 (0.11)	-0.19 (0.10)
Classroom Organization (Items 5,6,7)	0.11 (0.07)	0.08 (0.10)	0.10 (0.11)	0.13 (0.09)
Instructional Support (Items 8,9,10)	-0.10 (0.06)	-0.07 (0.09)	0.04 (0.11)	-0.07 (0.08)
Raw Average (3d Alternative) [a]				
Climate & Management (Items 1,2,5,6)	0.11 (0.07)	0.06 (0.11)	-0.01 (0.13)	0.11 (0.10)
Sensitivity & Regard (Items 3,4,7)	0.00 (0.08)	-0.03 (0.10)	-0.04 (0.13)	-0.16 (0.11)
Raw Average (2d Alternative) [a]				
Combine ES & CO (Items 1-7)	0.10 (0.06)	0.03 (0.08)	-0.05 (0.09)	-0.05 (0.08)
Confirmatory Bifactor				
General	0.03 (0.05)	0.00 (0.07)	-0.03 (0.07)	-0.09 (0.06)
Proactive Mgt. & Routines (Items 1-7)	-0.09 (0.05)	-0.05 (0.07)	-0.03 (0.08)	-0.13 (0.07)
Cognitive Facilitation (Items 8-10)	**-0.11 (0.05)***	-0.07 (0.07)	0.07 (0.11)	-0.02 (0.07)
FACES 2014				
	ROWPVT	Woodcock-Muñoz		
		Applied Problems	Letter Word	Spelling
Total Raw Score	-0.07 (0.08)	0.08 (0.09)	-0.05 (0.09)	0.05 (0.09)
Raw Average (3d Standard)				
Emotional Support (Items 1,2,3,4)	0.05 (0.13)	-0.02 (0.15)	-0.19 (0.19)	-0.09 (0.17)
Classroom Organization (Items 5,6,7)	-0.03 (0.15)	0.16 (0.19)	0.17 (0.22)	0.10 (0.23)
Instructional Support (Items 8,9,10)	-0.11 (0.08)	-0.03 (0.10)	-0.03 (0.10)	0.05 (0.11)
Raw Average (3d Alternative) [a]				
Climate & Management (Items 1,2,5,6)	-0.13 (0.09)	0.08 (0.17)	-0.13 (0.18)	-0.09 (0.16)
Sensitivity & Regard (Items 3,4,7)	0.16 (0.12)	0.03 (0.19)	0.08 (0.20)	0.07 (0.16)
Raw Average (2d Alternative) [a]				
Combine ES & CO (Items 1-7)	0.03 (0.07)	0.11 (0.10)	-0.05 (0.11)	-0.01 (0.10)
Confirmatory Bifactor				
General	0.02 (0.09)	0.12 (0.13)	0.06 (0.11)	0.06 (0.11)
Proactive Mgt. & Routines (Items 1-7)	-0.02 (0.07)	0.00 (0.10)	-0.14 (0.11)	-0.08 (0.10)
Cognitive Facilitation (Items 8-10)	-0.12 (0.07)	-0.01 (0.10)	-0.07 (0.10)	0.06 (0.10)

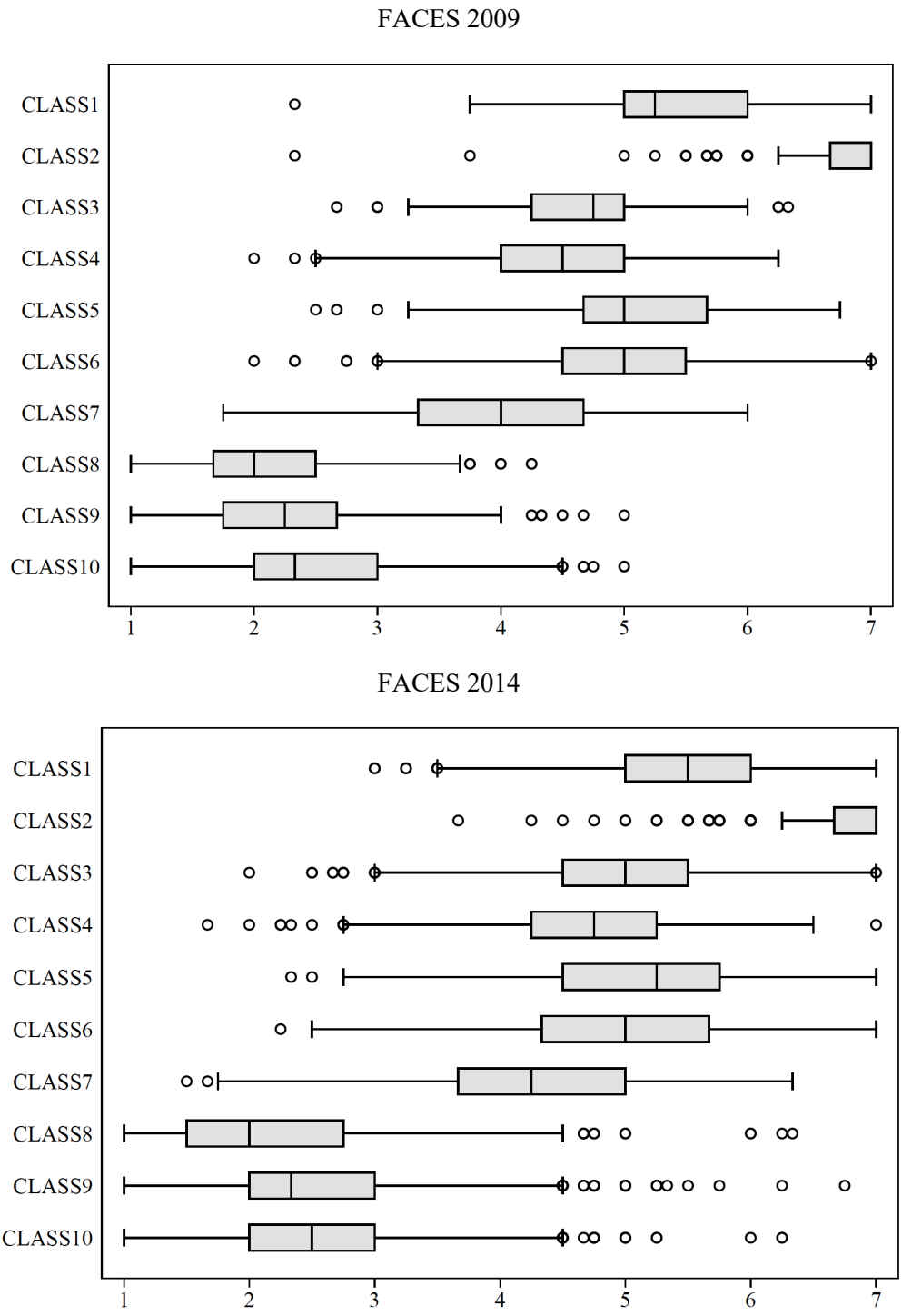
Note. Values are standardized coefficients (and standard errors) from five mixed models regressing children's spring scores on their classroom's spring CLASS score(s). The five models reflect the total score and four sets of scores. Covariates included the children's fall scores on the respective outcome as well as their gender and race-ethnicity, low-income, and disability status, whether their mother had a high school degree or less, their fall age cohort (3- or 4-year old), their age in months at the time of the spring assessment, and the months elapsed between the fall and spring assessments. The models included random intercepts at the classroom level and applied the child weights at the child-level and the classroom weights at the classroom level using the size approach to scaling the multi-level weights. Item-level missing values were imputed using multiple imputation (total *n* children = 397 and *n* classrooms = 140 in 2009; *n* children = 141 and *n* classrooms = 63 in 2014). * $p < .05$ (also bolded). [a] Results for the average raw score of Items 8-10 provided once, in the row labelled Instructional Support, to avoid duplication. ES = Emotional Support. CO = Classroom Organization. Mgt. = Management.

Table 6
Standardized Regression Coefficients for Child Social-Emotional Outcomes

	FACES 2009			
	Pencil Tapping	Leiter	Teacher-Reported	
CLASS Scores	Inhibitory Control	Attention/Social	Social Skills	Behavior Problems
Total Raw Score	0.04 (0.03)	0.02 (0.03)	0.05 (0.03)	-0.02 (0.03)
Raw Average (3d Standard)				
Emotional Support (Items 1,2,3,4)	-0.05 (0.04)	-0.04 (0.05)	-0.05 (0.05)	0.00 (0.04)
Classroom Organization (Items 5,6,7)	0.03 (0.05)	0.06 (0.04)	**0.11 (0.05)***	-0.05 (0.04)
Instructional Support (Items 8,9,10)	**0.08 (0.04)***	0.02 (0.03)	0.01 (0.04)	0.02 (0.03)
Raw Average (3d Alternative) [a]				
Climate & Management (Items 1,2,5,6)	0.06 (0.04)	0.02 (0.04)	-0.03 (0.05)	-0.02 (0.04)
Sensitivity & Regard (Items 3,4,7)	-0.08 (0.05)	-0.01 (0.05)	0.08 (0.05)	-0.02 (0.04)
Raw Average (2d Alternative) [a]				
Combine ES & CO (Items 1-7)	-0.02 (0.04)	0.01 (0.04)	0.05 (0.04)	-0.04 (0.03)
Confirmatory Bifactor				
General	0.02 (0.03)	0.02 (0.03)	0.05 (0.04)	-0.03 (0.03)
Proactive Mgt. & Routines (Items 1-7)	**-0.07 (0.03)***	-0.01 (0.03)	0.01 (0.03)	0.02 (0.03)
Cognitive Facilitation (Items 8-10)	**0.07 (0.03)***	0.01 (0.03)	0.00 (0.03)	0.01 (0.03)
	FACES 2014			
	Pencil Tapping	Leiter	Teacher-Reported	
	Inhibitory Control	Attention/Social	Social Skills	Behavior Problems
Total Raw Score	0.01 (0.06)	0.11 (0.06)	-0.02 (0.07)	0.04 (0.03)
Raw Average (3d Standard)				
Emotional Support (Items 1,2,3,4)	0.05 (0.09)	0.13 (0.10)	-0.13 (0.12)	-0.04 (0.06)
Classroom Organization (Items 5,6,7)	-0.09 (0.10)	-0.03 (0.11)	0.14 (0.14)	0.03 (0.08)
Instructional Support (Items 8,9,10)	0.05 (0.05)	0.03 (0.06)	-0.04 (0.06)	0.05 (0.04)
Raw Average (3d Alternative) [a]				
Climate & Management (Items 1,2,5,6)	-0.11 (0.09)	0.11 (0.10)	-0.12 (0.11)	0.10 (0.05)
Sensitivity & Regard (Items 3,4,7)	0.08 (0.10)	-0.01 (0.10)	0.13 (0.10)	**-0.11 (0.05)***
Raw Average (2d Alternative) [a]				
Combine ES & CO (Items 1-7)	-0.02 (0.07)	0.10 (0.06)	0.00 (0.07)	-0.01 (0.04)
Confirmatory Bifactor				
General	-0.01 (0.06)	0.02 (0.06)	0.01 (0.07)	0.01 (0.04)
Proactive Mgt. & Routines (Items 1-7)	-0.04 (0.05)	**0.14 (0.06)***	-0.02 (0.07)	0.00 (0.03)
Cognitive Facilitation (Items 8-10)	0.05 (0.05)	0.08 (0.05)	-0.01 (0.05)	0.03 (0.03)

Note. Values are standardized coefficients (and standard errors) from five mixed models regressing children's spring scores on their classroom's spring CLASS score(s). The five models reflect the total score and four sets of scores. Covariates included the children's fall scores on the respective outcome as well as their gender and race-ethnicity, low-income, and disability status, whether their academic assessment was in English, whether their mother had a high school degree or less, their fall age cohort (3- or 4-year old), their age in months at the time of the spring assessment, and the months elapsed between the fall and spring assessments. The models included random intercepts at the classroom level and applied the child weights at the child-level and the classroom weights at the classroom level using the size approach to scaling the multi-level weights. Item-level missing values were imputed using multiple imputation (total *n* children = 2,381 and *n* classrooms = 369 in 2009; *n* children = 974 and *n* classrooms = 193 in 2014, except for Pencil Tapping which was administered only to 4-year-olds and had *n*s of 999 and 268 in 2009 and 370 and 127 in 2014). * $p < .05$ (also bolded). [a] Results for the average raw score of Items 8-10 provided once, in the row labelled Instructional Support, to avoid duplication. ES = Emotional Support. CO = Classroom Organization. Mgt. = Management.

Figure 1
Box Plots of CLASS Items



Note. n = 370 (2009) and 641 (2014) classrooms. See Table 1 for item descriptions. CLASS2 reflects our reversal of scores for the negatively oriented negative climate item. Box plots are weighted by the classroom-level sampling weight.