AERA Open
January-December 2021, Vol. 7, No. 1, pp. 1–19
DOI: 10.1177/23328584211038938
Article reuse guidelines: sagepub.com/journals-permissions

Variability in Preschool CLASS Scores and Children's School Readiness

Jennifer K. Finders
Adassa Budrevich
Robert J. Duncan
David J. Purpura
James Elicker
Sara A. Schmitt
Purdue University

The Classroom Assessment Scoring System (CLASS) is a widely administered measure of classroom quality that assesses teacher-child interactions in the domains of Emotional Support, Classroom Organization, and Instructional Support. We use data from an evaluation of state-funded prekindergarten provided to 684 children from families with low incomes (M_{age} = 57.56 months, 48% female) to examine the extent to which CLASS scores vary over the course of an observational period within a single day and investigate whether this variability is related to children's school readiness at the end of the preschool year, holding constant two additional measures of quality. Teacher-child interactions in all three domains were moderately stable. Mean Classroom Organization was positively related to math, and variability in Classroom Organization was negatively related to literacy. Mean Instructional Support was negatively associated with language. Findings have implications for programs that utilize the CLASS to make high-stakes decision and inform professional development.

Keywords: *classroom quality, preschool, teacher-child interactions, school readiness, educational policy, evaluation*

Income disparities in children's school readiness skills are key contributors to pervasive achievement gaps (Burchinal et al., 2011; Durham et al., 2007; von Hippel et al., 2018). Policy efforts concerning early childhood education (ECE), such as increasing access to state-funded prekindergarten programs implemented within Quality Rating and Improvement Systems (QRIS), have been initiated to support school readiness skills and close achievement gaps (Barnett et al., 2018). One specific feature of such programs that is intended to promote children's learning is high-quality teacher-child interactions (Sabol & Pianta, 2015). However, effects of teacher-child interactions on children's outcomes, as measured by the widely used Classroom Assessment Scoring System (CLASS), are consistently small (Keys et al., 2013; Perlman et al., 2016) or null (Guerrero-Rosada et al., 2021). Recent work has begun to reconsider the ways in which CLASS scores are utilized in analyses (e.g., Hatfield et al., 2016). For instance, researchers argue that scores in the mid-range of the CLASS may not be picking up on meaningful differences in the quality of children's experiences (Burchinal, 2018), and capturing subtle inconsistencies in teacher-child interactions over the course of a day may be a more sensitive indicator than average levels of quality (Brock et al., 2018; Curby et al., 2013). Yet the consequences of variable teaching practices during the preschool year have not been fully explored across CLASS domains and school readiness outcomes, leaving unanswered questions as to how a predominant measure of quality can be used most effectively to inform professional development and policy. The present study begins to fill this gap by analyzing the links between variability in the CLASS domains of Emotional Support, Classroom Organization, and Instructional Support and growth in literacy, language, and math in preschool among a sample of children from families with low incomes. Importantly, analyses control for two additional measures of quality (mean CLASS scores and state QRIS scores) to determine whether variability emerges as a distinct indicator of quality and predictor of outcomes.

Preschool Classroom Quality and the CLASS

Classroom quality in ECE settings is often conceptualized in terms of structure-oriented indicators (e.g., class size, teacher education) and processes-oriented features (e.g., warmth, stimulation) that create optimal learning opportunities (Burchinal et al., 2014). Although both are important, classroom processes serve as the primary mechanisms of development and learning (Mashburn et al., 2008). Specifically, effective teacher-child interactions have been identified as a key ingredient of process quality that

promotes school readiness (Hamre & Pianta, 2007; Hong et al., 2019). Moreover, because observational measures of process quality have been recognized as an essential component of ECE quality, they are increasingly becoming integrated into state-funded prekindergarten programs (Friedman-Krauss et al., 2020). Currently, the CLASS is the dominant observation system used in research, practice, and policy to assess process quality. It is based on theory suggesting that high-quality interactions support development through providing children with a sense of security and feelings of connectedness (Ainsworth, 1989; Hamre & Pianta, 2001), ensuring behaviors and instructional activities are well managed (Cameron et al., 2005), and scaffolding learning and understanding of complex concepts (Davis & Miyake, 2004; Vygotsky, 1991). The CLASS measures three different aspects of teacher-child interactions (Pianta et al., 2008). *Emotional Support* indicates the degree of warmth and positivity within the classroom as well as the teacher's responsiveness and regard for student perspectives. *Classroom Organization* denotes the effectiveness of behavior management strategies, provision of activities, and overall productivity within the classroom. *Instructional Support* is represented by the quality of instruction and feedback provided by the teacher.

Theoretical work emphasizes connections between the CLASS domains and children's school readiness (Downer et al., 2010). Although some empirical studies have documented relations between high-quality teacher-child interactions and the development of school readiness using the CLASS (e.g., Araujo et al., 2016; Burchinal et al., 2008; Hamre et al., 2014; Howes et al., 2008; Leyva et al., 2015), replication work and meta-analyses reveal null or weak associations between mean CLASS scores and child outcomes, with effect sizes falling between .04 and .09 (Guerrero-Rosada et al., 2021; Keys et al., 2013; Perlman et al., 2016; Weiland et al., 2013). Further, data suggest that only a small percentage of children actually experience high-quality instruction. For instance, one study revealed that between 76% and 87% of classrooms failed to reach the threshold of classroom quality necessary for supporting school readiness development (Hatfield et al., 2016), suggesting that teachers are unlikely to maintain consistent, high-quality interactions with children throughout the day. Indeed, most CLASS scores tend to fall within the low- to mid-range of the distribution for Classroom Organization and Instructional Support, indicating that teachers fluctuate in their provision of high- and low-quality behaviors (La Paro et al., 2009). Together, these factors may be limiting the predictive validity of the CLASS for child outcomes.

Reconsidering the ways in which CLASS scores are conceptually and methodologically used in research and practice may help overcome these limitations. The CLASS procedure involves several 20-minute observations of teacher-child interactions within the same school day, and each observation is followed by 10-minutes of coding. Scores for each interval are assigned along a 7-point scale, with 1 or 2 indicating low quality, 3 to 5 indicating mid-range quality, and 6 or 7 indicating high quality. Typically, CLASS scores are averaged across these intervals for each domain (Emotional Support, Classroom Organization, and Instructional Support), eliminating variability between the intervals. These averages have been used to determine certain thresholds that are necessary to promote development (Burchinal et al., 2010; Hatfield et al., 2016), which have subsequently been incorporated into state and federal initiatives, including Head Start (U.S. Department of Health and Human Services, 2020). However, the relatively weak effect sizes documented in the extant literature has led many to question whether the use of the CLASS for these purposes is appropriate (Gordon & Peng, 2020; Mantzicopoulos et al., 2018; Mashburn, 2017). In addition to using mean scores, researchers argue that there may be meaningful variability captured in observational measures of classroom quality that are obscured with averages (e.g., Curby et al., 2013; LoCasale-Crouch et al., 2018).

Rationale for Examining Variability in CLASS Scores

Children's development is shaped by their dynamic social interactions with adults—typically referred to as proximal processes (Bronfenbrenner & Morris, 2016). Adult caregivers help children make sense of the world through engaging in warm, supportive, and stimulating interactions, which promote children's sense of security and foster a healthy attachment (Ainsworth et al., 1978). An important but often implicit feature of this model is the notion that interactions follow a predictable pattern to allow children to anticipate how their needs may be met by their caregiver. Thus, by definition, proximal processes need to be consistent in order to be effective (Bronfenbrenner & Morris, 1998). Extended to adults within the classroom, inconsistencies in the moment-to-moment interactions that children have with their teachers may create issues of mistrust, confusion, and missed opportunities, which could ultimately affect children's development (Curby et al., 2009). For instance, children may be hesitant to participate in learning activities and take risks if they fear their teacher may react unpredictably. They may become distracted from a particular goal if their teacher is disorganized and lacks the ability to effectively facilitate learning activities. Finally, children may be unable to learn complex concepts and vocabulary if their teacher does not consistently scaffold at their level. These examples demonstrate how variability in teacher-child interactions may undermine children's development through their withdrawal in learning (Vitiello et al., 2012). This type of variability is thought to be especially detrimental to children who are from families with low incomes (Tran & Winsler, 2011), perhaps because they are less likely to experience

high-quality teacher-child interactions over the first few years of schooling (Pianta et al., 2007).

In addition to the conceptual argument, there are also methodological advantages to considering variability as an indicator of quality. It has been hypothesized that low levels of variability in mean scores on quality rating scales, particularly when it comes to distinguishing classrooms with scores in the mid-range, may be responsible for the underestimation of effects (Burchinal, 2018). In usual practice, collapsing CLASS scores across intervals to arrive at an average may mask meaningful differences between classrooms. Researchers have argued that the variability between observational intervals within a day may actually contain important information about children's learning environments (e.g., Curby et al., 2011). For instance, Snow and Matthews (2016) contend that "consistent feedback to and interaction with students" is an essential component of high-quality teaching that fosters language and literacy skills in preschool and the early grades (p. 69). Moreover, a fundamental assumption underlying the CLASS framework is that high-quality teacher-child interactions are not only nurturing and responsive but also consistent (Bailey et al., 2013). The intent to measure the degree of consistency in teacher behaviors is also evident in the coding scheme outlined by the developers of the CLASS in the training manual (Pianta et al., 2008). Thus, the CLASS in its design offers the opportunity to capture variability by taking into consideration scores at each interval of observation.

Sources of Variability in CLASS Scores and Children's Outcomes

Variability is conceptualized as fluctuations in the quality of teacher-child interactions that are due to systematic variation (Curby et al., 2011). A variety of methods have been implemented to measure such variability (Wang et al., 2020). The most commonly used technique is to quantify variability through the standard deviation (Curby et al., 2010). Yet another popular approach is to utilize generalizability theory (e.g., Mashburn et al., 2014; Praetorius et al., 2014). The goal of generalizability studies (G-studies) is to understand reliability and stability in ratings over time (Shavelson & Dempsey-Atwood, 1976). In essence, G-studies decompose variability in observational ratings of classroom quality into different components, their interactions, and measurement error. This partitioning of variance can inform decisions regarding the improvement of measures by helping to identify the optimal data collection strategy and scoring criteria for a desired reliability (Hill et al., 2012). For instance, results from one G-study yielded evidence that only 10% to 45% of the variance in kindergarten CLASS scores is attributed to the behaviors of teachers, with a large amount of variability not explained by teachers, lessons, or raters (Mantzicopoulos et al., 2018). These findings suggest that it may require more observations and raters to achieve accurate and stable ratings than is likely feasible in practice (Praetorius et al., 2014).

A growing body of literature has analyzed sources of variability in ratings of classroom quality as it relates to classroom schedules, routines, and activity types, using various methods. For instance, CLASS scores have been shown to vary by season (Buell et al., 2017), from lesson to lesson (Patrick & Mantzicopoulos, 2016), between subject areas (Cohen et al., 2018; Kook & Greenfield, 2020), by ECE program type (Bassok et al., 2021), and CLASS observations of shorter durations tend to be scored more favorably (Cash & Pianta, 2014). Additionally, researchers have documented significant variability in CLASS scores within a single day (Curby et al., 2010). Notably, CLASS scores within the domains of Classroom Organization and Instructional Support appear to decrease over the course of an observation period (Thorpe et al., 2020; Wang et al., 2020). Together, this research supports the existence of variability in teacher-child interactions over the day, some of which may be a function of classroom characteristics, such as teacher-to-child ratio and length of the school day (Von Suchodoletz et al., 2014), and some of which may be due to measurement "noise" (Casabianca et al., 2015). Evaluating whether variability is a distinct dimension of classroom quality that provides meaningful information about children's learning environment beyond the aggregate approach of averaging CLASS scores may help clarify if observed fluctuations in teacher-child interactions are in fact representative of lower overall classroom quality. Thus, in the present study, we utilize the full distribution of CLASS scores within an observation period to explore variability in classroom quality within a single day and links to school readiness.

Few studies have investigated the stability of teacher–child interactions with regard to children's academic trajectories, and evidence suggests there may be negative consequences of variability for children's learning (Cash et al., 2019). Prior work using the CLASS to analyze variability within a single day has focused almost entirely on the Emotional Support domain (e.g., Curby et al., 2013). Results indicate that consistency (i.e., lack of variability) in Emotional Support over the day is positively associated with children's language and rhyming skills, above-and-beyond mean level Emotional Support (Curby et al., 2013). In addition, Emotional Support consistency appears to be significantly related to fewer problem behaviors in preschool and kindergarten (Brock & Curby, 2014; Zinsser et al., 2013). In a recent study with an older sample, variability in Instructional Support was negatively related to math performance in elementary school (Sandilos et al., 2019). However, previous work has not investigated the effects of variability in classroom quality for all CLASS domains on school readiness indicators during the preschool year. Given the limited work in this area, it is impossible to draw conclusions about whether variability in CLASS scores can be considered a

distinct indicator of quality that provides meaningful information beyond the mean, primarily because these effects have not been replicated across domains or samples with a robust set of control variables (Duncan et al., 2014).

The Present Study

In the present study, we had two research aims. The first aim was to determine the extent to which classrooms varied in Emotional Support, Classroom Organization, and Instructional Support over four observation intervals within a single day using the standard deviation approach. Based on previous research, we expected that CLASS scores would be moderately stable across intervals within a 2-hour observation period (Curby et al., 2010), and that quality in Classroom Organization and Instructional Support would vary considerably more than quality in Emotional Support (Thorpe et al., 2020; Wang et al., 2020). The second aim was to investigate whether variability in all three CLASS domains emerged as a distinct and more robust indicator of school readiness than average CLASS scores while taking into account two additional measures of quality (mean CLASS scores and state QRIS scores). We hypothesized that experiencing volatile Emotional Support in the form of inconsistent warmth and responsiveness from teachers may inhibit children's ability to feel safe and take risks while participating in the learning process. A similar hypothesis was drawn for Classroom Organization, based on our assumption that children may disengage in learning when they are subjected to a chaotic environment where expectations are unclear. With regard to Instructional Support, we suspected that variability could actually promote development. Given that scores in this domain are typically low (e.g., Hamre, 2014; Hatfield et al., 2016), providing at least some mid- or high-quality interactions in otherwise rote instructional environments may be necessary for learning (Brock et al., 2018). To illustrate, interventions that have been successful in improving mean classroom quality tend to reduce variability in Emotional Support and Classroom Organization but increase variability in Instructional Support (Early et al., 2017). However, we left this hypothesis as exploratory based on the findings of one empirical study indicating the contrary (Sandilos et al., 2019).

Method

Participants

The sample for this study included three cohorts of children (N = 684; 48% female) across 180 preschool classrooms (M = 3.5 children per classroom) in 127 schools (M = 1.42 classrooms per school, M = 4.8 children per school) who participated in a larger study focused on evaluating the impacts of a state-funded prekindergarten program on children's school readiness. The sample was racially and ethnically diverse and represented the broader area, with most parents identifying their children as Black/African American (43%) or White/Caucasian (32%). Children were eligible to participate in the larger evaluation study if they were at least 4 years old at the start of the preschool year (M_{age} = 57.56 months; SD = 3.76 months) and if their family incomes fell at or below 127% of the federal poverty line. Teacher demographics are presented in Table A1 of the appendix.

The larger evaluation study utilized a quasi-experimental design to compare school readiness between children in the state-funded prekindergarten group (67% of sample) who attended high-quality preschools rated as Level 3 or 4 on the state's QRIS, and children in the comparison group who attended low-quality preschools rated as Level 1 or 2 (or not enrolled in the QRIS). Child care programs across eight counties in the state were invited to participate in the study if (1) they accepted child care development funds (CCDF) and their program was rated as Level 0, (not enrolled), 1, or 2 on the QRIS or (2) they were an approved state-funded prekindergarten provider rated as Level 3 or 4 on the QRIS. All parents who used CCDF vouchers in the comparison condition and all parents of children in the state-funded prekindergarten program were invited to participate. The analytic sample includes all children from the larger evaluation study.

Procedures

Trained research assistants administered direct assessments of literacy, vocabulary, and mathematics to children in the fall and spring, and parents filled out a demographic questionnaire in the fall of the preschool year. Classroom quality was observed by CLASS-certified research assistants during the winter of the preschool year. Families and teachers received a $20 compensation in the fall and spring for their participation.

Measures

School Readiness. Children's school readiness was assessed with three measures of early academic skills.

Literacy. Literacy was measured through the Letter-Word Identification subtest of the Woodcock Johnson–IV (WJLW-IV; Schrank et al., 2014). The WJLW-IV subtest requires children to use their receptive and expressive literacy skills as they identify letters and words. The subtest contains 76 items grouped into 15 sets. Children reach ceiling once they respond incorrectly to six consecutive items to finish out a set. Raw scores ranged from 0 to 47 in the spring of preschool. The WJLW-IV has a reliability of .84 to .94 for children ages 2 to 7 years (Villarreal, 2015).

Language. Children's receptive vocabulary was assessed via the Peabody Picture Vocabulary Test–fourth edition (PPVT-IV; Dunn & Dunn, 2007). Children are presented with four simultaneous images and are asked to point to a

picture that represents the verbal cue provided by the assessor. The PPVT-IV includes 228 items grouped into 13 sets. Children must respond correctly to all but one item in the first set before moving forward with the task. Children reach the ceiling once they respond incorrectly to eight items incorrectly in a set. Raw scores ranged from 14 to 144 in the spring of preschool. The PPVT-IV has strong internal consistency ($\alpha = .94$; Dunn & Dunn, 2007).

Math. Children's math was assessed with the Applied Problems subtest of the Woodcock Johnson–IV (WJAP-IV; Schrank et al., 2014). The WJAP-IV subtest assesses quantitative knowledge and reasoning by requiring children to solve orally presented math problems. The subtest contains 55 items grouped into 14 sets. Children reach the ceiling once they respond incorrectly to five consecutive items to finish out a set. Raw scores ranged from 0 to 21 in the spring of preschool. The WJAP-IV has also demonstrated high reliability (Villarreal, 2015).

Preschool Classroom Quality. Classroom observations using the Pre-K CLASS (Pianta et al., 2008) were conducted to measure the quality of teacher-child interactions. The CLASS is composed of three domains that contain multiple dimensions: Emotional Support (positive climate, negative climate, teacher sensitivity, and regard for students perspectives), Classroom Organization (behavior management, productivity, and instructional learning formats), and Instructional Support (concept development, quality of feedback, and language modeling). Research assistants completed a two-day Pre-K CLASS training provided by Teachstone. After the training they were required to pass the CLASS reliability test and score within 80% of the master codes across five videos in order to become a certified CLASS observer. Certified CLASS observers rated classrooms on each of the dimensions using a 7-point scale (1 or 2 = *low quality*, 3 to 5 = *mid quality*, 6 or 7 = *high quality*) over four 20-minute intervals with 10 minutes of coding after each. These cycles took place over the course of roughly 2 hours in a single preschool day, either in the morning or afternoon, during the winter months (e.g., January through March).

Mean CLASS. Mean classroom quality for Emotional Support, Classroom Organization, and Instructional Support was calculated by first averaging each set of dimensions for a particular domain within intervals to arrive at an interval-specific domain score, and then averaging the interval-specific domain scores across the four intervals. The internal consistency for all three CLASS domain scores within the current sample was high: Emotional Support ($\alpha = .93$), Classroom Organization ($\alpha = .90$), and Instructional Support ($\alpha = .83$). Previous findings show that mean scores on the CLASS are moderately correlated with the Early Childhood Environmental Rating Scale–Revised ($r = .52$ for Emotional Support, $r = .40$ for Instructional Support; La Paro et al., 2004), which is another widely used global scale for classroom quality in the field (Harms et al., 1998).

Variability in CLASS. Variability in classroom quality in Emotional Support, Classroom Organization, and Instructional Support was represented by the standard deviation between the four observation intervals. We calculated the average variance across intervals by subtracting each of the four interval-specific domain scores from the mean within a single CLASS domain, squaring the resulting values to put convert them to a positive scale, adding them together, and dividing by the number of observations minus one for each classroom (i.e., $n - 1$). The resulting value represented the average amount of variability within a classroom around the mean. This approach has the advantage of capturing the entire spread of an individual classroom's score (Curby et al., 2013) and has been shown to produce similar estimates to other statistical methods of calculating variability (Wang et al., 2020).

State QRIS Scores. The state QRIS contains four levels that each build on the foundation of the previous level, resulting in significant quality improvements at each stage. High-quality programs are those that receive Level 3 and 4 ratings. The general criteria for achieving each level of quality within the QRIS are as follows: (1) health and safety needs of children are met, (2) environment supports children's learning, (3) planned curriculum guides child development and school readiness, and (4) national accreditation is achieved. It is important to note, however, that the state QRIS does not incorporate CLASS scores to determine a program's quality level. In previous research, scores on the state QRIS have shown modest associations with global classroom quality (Elicker et al., 2011; Lahti et al., 2015). We created a categorical variable to represent the state's definition of high-quality (QRIS = 3 or 4), low-quality (QRIS = 1 or 2), and programs that were unrated (QRIS = 0) for analyses.

Covariates. Child age and sex (1 = female, 0 = male) were included as covariates because of the extant research demonstrating that older children and girls are more likely to have higher school readiness in preschool (Bornstein et al., 1998; Song et al., 2015). We also controlled for cohort, teacher-to-child ratio, teacher education level, teacher experience, and whether or not the teacher had a Child Development Associate (CDA) credential because of their potential influence on CLASS scores and the outcomes of interest (Von Suchodoletz et al., 2014).

Analytic Plan

To explore the extent to which classroom quality varied within a single day, we first ran unconditional multilevel

models with CLASS scores at each of the four intervals nested within classrooms using the MIXED command in Stata 14 (StataCorp, 2015). The intraclass correlations (ICCs) from the unconditional models provided an indication of how correlated each interval was with each other over time, with a lower ICC signifying less consistency (i.e., greater variability) in classroom quality over the observation period within a day. Next, to examine whether variability in CLASS scores was a unique and more robust indicator of quality, we utilized structural equation modeling (SEM) to estimate a series of nested regression models that tested whether variability in classroom quality, computed as the average variance, significantly predicted continuous scores on school readiness outcomes holding constant child- and classroom-level covariates. In all models, mean CLASS scores and state QRIS scores were also included as covariates in order to investigate whether modeling variability contributed substantial information about children's development above and beyond typical aggregated measures of classroom quality.

The ICCs for the school readiness outcomes at the classroom-level were high for literacy (.29), math (.16), and vocabulary (.15). Therefore, we clustered the standard errors within classrooms to handle the nonindependence of data. When the CLUSTER option is specified with SEM, Stata produces almost identical point estimates and standard errors as multilevel models using the MIXED command (Stapleton et al., 2016). To contextualize results within the broader field, which typically utilizes CLASS averages, we ran regression models in a stepwise fashion adding in the following covariates at each iteration: (1) child- and classroom-level covariates, (2) mean CLASS scores, and (3) variability in CLASS scores. By examining the pseudo R^2, we isolated the amount of variance explained by each of the target variables. Regression models were estimated separately for each CLASS domain (Emotional Support, Classroom Organization, and Instructional Support), but outcomes and the covariance between outcomes were modeled simultaneously within the SEMs. The battery of child- and classroom-level covariates included all the aforementioned variables as well as children's baseline skills at the fall of preschool, which enabled us to investigate residualized change in children's school readiness at the end of preschool.

Missing Data

There was a small amount of missingness on the primary variables of interest. Only six classrooms (5%) were missing CLASS scores and four classrooms (4%) were missing state QRIS scores. Between 27 and 48 children (4%–7%) were missing data on fall school readiness assessments and between 82 and 97 children (12%–14%) were missing data on spring school readiness assessments. There was very little missing data on child covariates (<1%), but a fair amount of missing data on teacher-to-child ratio (23%) as well as teacher education and experience, and whether teachers had their CDA (56%–58%). Full-information maximum likelihood was used to account for missing data. It produces estimates that are less biased than listwise deletion and allows for all observations to inform model estimates (Acock, 2012).

Results

Descriptive statistics and bivariate correlations for primary study variables are presented in Table 1.

Variability in CLASS Scores Within a Day

The first research aim examined the extent to which preschool CLASS scores varied during the observation period within a day. Means and standard deviations for Emotional Support, Classroom Organization, and Instructional Support at each of the four intervals indicated that classroom quality in all three domains decreased over the observation period within a day (Table 2). ICCs for the unconditional multilevel models at the within-classroom level were highest for Emotional Support (.80), followed by Classroom Organization (.72) and Instructional Support (.66), suggesting that CLASS scores at each interval were moderately correlated with each other over time. The lower ICC for Instructional Support indicated that quality in this domain was less consistent (i.e., more variable) than quality in Emotional Support and Classroom Organization during the day. Although there was a fair degree of stability in CLASS scores within classrooms, the remaining variance implied that fluctuations in quality that could be subsequently examined.

Variability in CLASS Scores and Children's School Readiness

The second research aim investigated whether variability in CLASS scores during an observation period was related to children's school readiness, after controlling for mean CLASS scores and state QRIS scores. Results from SEMs revealed that neither mean Emotional Support nor variability in Emotional Support were significantly related to any of the school readiness outcomes (Table 3). Mean Classroom Organization positively predicted spring math after accounting for child- and classroom-level covariates ($b = 0.20$, $SE = 0.10$, $p = .04$, B = 0.07), such that children in classrooms that were generally productive and well managed had greater growth in math skills from fall to spring of the preschool year (Table 4). However, this association became only marginally significant after including variability in Classroom Organization in the model ($b = 0.19$, $SE = 0.10$, $p = .06$, B = 0.06). Additionally, variability in Classroom Organization negatively predicted children's literacy during the preschool

TABLE 1
Descriptive Statistics and Correlations

Variable	1	2	3	4	5	6	7	8	9	*M*	*SD*	Range
1. Mean ES	—									5.24	1.00	2.81–6.88
2. Mean CO	.74***	—								4.58	1.14	1.00–6.75
3. Mean IS	.50***	.51***	—							2.22	0.91	1.00–5.25
4. ES Variability	−.33***	−.27***	−.15***	—						0.31	0.39	0.00–3.02
5. CO Variability	−.04	−.18***	.02	.52***	—					0.62	0.73	0.00–4.62
6. IS Variability	.20***	.16***	.42***	.32***	.37***	—				0.56	0.75	0.00–5.44
7. Spring language	.03	.11*	−.08	−.08	−.08	−.13**	—			77.89	21.50	14.0–144.0
8. Spring literacy	−.07	.12**	−.03	−.09*	−.14***	−.13**	.38***	—		9.92	5.50	0.00–47.0
9. Spring math	−.05	.05	−.09*	−.03	−.05	−.07	.59***	.45***	—	11.60	3.52	0.00–21.0
10. QRIS score	.23***	.06	.15***	−.10*	−.02	.07	.02	−.16***	−.03	1.50	0.76	0.00–2.00

Note. ES = Emotional Support; CO = Classroom Organization; IS = Instructional Support; QRIS = Quality Rating and Improvement System (0 = not rated, 1 = low-quality, 2 =high quality).
*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 2
Interval-Specific Means, Standard Deviations, and Correlations for Classroom Quality by Domain (n = 180 Classrooms)

Classroom quality domain	Interval 1, *M* (*SD*)	Interval 2, *M* (*SD*)	Interval 3, *M* (*SD*)	Interval 4, *M* (*SD*)	ICC, *r*
Emotional Support	5.40 (1.09)	5.22 (1.12)	5.21 (1.15)	5.09 (1.15)	0.80
Classroom Organization	4.77 (1.35)	4.50 (1.30)	4.44 (1.35)	4.35 (1.32)	0.72
Instructional Support	2.38 (1.28)	2.20 (1.15)	2.18 (1.13)	2.02 (1.05)	0.66

Note. ICC = intraclass correlation.

year ($b = -0.65$, $SE = 0.17$, $p < .001$, B = −0.09), holding mean Classroom Organization and the battery of child- and classroom-level covariates constant. Children demonstrated less growth in literacy skills from fall to spring of preschool when teachers were more variable, or less consistent, in setting expectations and engaging in behavior management strategies during the observation period (Table 4). However, mean Classroom Organization was not significantly related to children's literacy skills. Furthermore, neither variability in Classroom Organization nor mean Classroom Organization were significantly related to children's language skills (Table 4). Finally, mean Instructional Support negatively predicted children's language skills both on its own ($b = -1.75$, $SE = 0.59$, $p = .003$, B = −0.07) and when included in the model with variability in Instructional Support ($b = -1.85$, $SE = 0.74$, $p = .01$, B = −0.08), holding child- and classroom-level covariates constant. Children demonstrated less growth in language from fall to spring of preschool when teachers engaged in more frequent conversations that required higher-level thinking. However, variability in Instructional Support was not significantly related to children's language skills. Additionally, neither variability in Instructional Support nor mean Instructional Support were significantly related to children's literacy or math outcomes (Table 5).

Discussion

In the present study, we examined whether variability in preschool classroom quality during an observation period within a single day emerged as a unique and more robust indicator of children's school readiness than typical aggregate measures. Using the CLASS, we were able to explore variability in teacher–child interactions across the domains of Emotional Support, Classroom Organization, and Instructional Support. Results from unconditional multilevel models with CLASS observations nested within classrooms revealed that there was moderate to high consistency (i.e., low variability) in quality over four intervals of classroom observation. These findings align with previous work documenting relative stability in Emotional Support (*r*s = .64–.77), Classroom Organization (*r*s = .55–.71), and Instructional Support (*r*s = .52–.64; Curby et al., 2010; Curby et al., 2011). Moreover, our results are congruent with recent work demonstrating lower consistency in Instructional Support and Classroom Organization relative to Emotional Support when investigating several methods of measuring variability (Wang et al., 2020). It should be noted that quality in all three domains decreased over the observation period. Research suggests that children's positive engagement with teachers and peers may taper off over the preschool day (Vitiello et al., 2012). Moreover, decreasing quality over the course of a 2-hour observation may be explained by the changing contexts of the preschool classroom (Thorpe et al., 2020). During free-play and large group instruction, teachers may remain responsive and use linguistically rich language relative to mealtimes and routines, in part, because they are able to engage with the whole classroom (Cabell et al., 2013; Turnbull et al., 2009). Thus, it is possible that teachers face the challenge of eliciting high-quality interactions as their energy depletes and children simultaneously lose focus or interest in activities over the progression of the day.

With regard to variability in CLASS scores and children's learning, results revealed that variability in Emotional Support was not significantly related to children's growth in school readiness during the preschool year. Previous research has uncovered links between variability in Emotional Support and expressive language skills (Curby et al., 2013). These authors suggest that inconsistencies in the provision of Emotional Support may indicate volatile relationships, which in turn, discourage children's expressiveness. Although inconsistent with Curby et al. (2013), our results do align with one recent study that also did not find significant associations between variability in Emotional Support and academic outcomes in later grades (Sandilos et al., 2019). The null findings in the present study may be explained by the fact that Emotional Support was the most consistent domain of quality with the highest mean, implying there was very little within- and between-classroom variability in Emotional Support. A relatively restricted range of Emotional Support has also been documented in previous research (e.g., Qi et al., 2019). Moreover, mean Emotional Support was not significantly related to any school readiness outcomes. These findings indicate that the way Emotional Support quality is measured may not always capture the interactions that are theorized to support children's development, particularly in preschool classrooms where teachers

TABLE 3

Results From Regression Models With Emotional Support Predicting School Readiness Outcomes

	Emotional Support								
	Literacy			Language			Math		
Variable	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Child-level controls									
Age in months	0.03 (0.04)	0.03 (0.04)	0.03 (0.04)	0.45** (0.13)	0.47** (0.13)	0.47*** (0.13)	0.06* (0.03)	0.07* (0.03)	0.07* (0.03)
Female	0.19 (0.27)	0.21 (0.27)	0.20 (0.27)	0.34 (0.96)	0.40 (0.96)	0.38 (0.97)	−0.13 (0.22)	−0.13 (0.22)	—0.13 (0.22)
Cohort	−0.64** (0.23)	−0.61** (0.23)	−0.56* (0.23)	−0.33 (0.86)	−0.26 (0.92)	−0.29 (0.98)	−0.26 (0.19)	−0.27 (0.19)	−0.30 (0.20)
Fall autoregressor	0.93*** (0.06)	0.93*** (0.06)	0.93*** (0.06)	0.82*** (0.03)	0.82*** (0.03)	0.82*** (0.03)	0.63*** (0.03)	0.62*** (0.03)	0.62*** (0.03)
Classroom-level controls									
Teacher–child ratio	−4.35 (3.18)	−3.60 (3.19)	−2.96 (3.08)	−14.28 (10.68)	−12.92 (10.64)	−12.17 (11.10)	−3.43 (2.25)	−3.43 (2.24)	−3.49 (2.28)
Teacher education	0.21 (0.17)	0.20 (0.18)	0.19 (0.18)	0.71 (0.82)	0.73 (0.85)	0.71 (0.84)	−0.35* (0.15)	−0.35* (0.16)	−0.35* (0.16)
Teacher CDA credential	−0.71 (0.69)	−0.68 (0.70)	−0.71 (0.68)	−1.77 (2.97)	−1.62 (2.96)	−1.55 (2.95)	0.38 (0.42)	0.38 (0.44)	0.41 (0.42)
Teacher experience	0.03 (0.03)	0.03 (0.03)	0.02 (0.03)	−0.08 (0.19)	−0.08 (0.20)	−0.11 (0.20)	−0.01 (0.03)	−0.01 (0.03)	−0.01 (0.03)
QRIS level	−0.13 (0.29)	−0.05 (0.29)	−0.05 (0.29)	1.13 (0.93)	1.31 (0.92)	1.34 (0.92)	0.04 (0.16)	0.02 (0.17)	0.02 (0.17)
Classroom-level quality									
CLASS mean		−0.28 (0.20)	−0.34† (0.20)		−0.73 (0.76)	−0.82 (0.74)		0.01 (0.15)	0.01 (0.15)
CLASS variability			−0.69† (0.39)			−0.87 (2.02)			0.07 (0.29)
Pseudo R^2	.63	.63	.64	.68	.68	.68	.52	.52	.52

Note. Unstandardized betas reported; standard errors in parentheses. CLASS = Classroom Assessment Scoring System; QRIS = Quality Rating and Improvement System; CDA = Child Development Associate.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 4
Results From Regression Models With Classroom Organization Predicting School Readiness Outcomes

	Classroom Organization								
	Literacy			Language			Math		
Variable	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Child-level controls									
Age in months	0.03 (0.04)	0.02 (0.04)	0.03 (0.04)	0.45** (0.13)	0.46** (0.13)	0.47*** (0.13)	0.06* (0.03)	0.07* (0.03)	0.07* (0.03)
Female	0.19 (0.27)	0.17 (0.26)	0.19 (0.26)	0.34 (0.96)	0.35 (0.97)	0.42 (0.97)	−0.13 (0.22)	−0.15 (0.21)	−0.15 (0.21)
Cohort	−0.64** (0.23)	−0.72** (0.22)	−0.72** (0.21)	−0.33 (0.86)	−0.46 (0.95)	−0.44 (0.94)	−0.26 (0.19)	−0.30 (0.18)	−0.29 (0.18)
Fall autoregressor	0.93*** (0.06)	0.93*** (0.06)	0.92*** (0.06)	0.82*** (0.03)	0.82*** (0.03)	0.81*** (0.03)	0.63*** (0.03)	0.61*** (0.03)	0.61*** (0.03)
Classroom-level controls									
Teacher–child ratio	−4.35 (3.18)	−4.33 (3.21)	−3.69 (2.97)	−14.28 (10.68)	−14.25 (10.78)	−12.85 (10.58)	−3.43 (2.25)	−3.44 (2.28)	−3.32 (2.35)
Teacher education	0.21 (0.17)	0.11 (0.18)	0.07 (0.17)	0.71 (0.82)	0.52 (0.98)	0.48 (0.96)	−0.35* (0.15)	−0.44** (0.15)	−0.44** (0.16)
Teacher CDA credential	−0.71 (0.69)	−0.65 (0.71)	−0.72 (0.64)	−1.77 (2.97)	−1.54 (3.02)	−1.64 (2.95)	0.38 (0.42)	0.36 (0.40)	0.34 (0.39)
Teacher experience	0.03 (0.03)	0.02 (0.03)	0.01 (0.03)	−0.08 (0.19)	−0.10 (0.20)	−0.11 (0.20)	−0.01 (0.03)	−0.02 (0.03)	−0.02 (0.03)
QRIS level	−0.13 (0.29)	−0.08 (0.30)	−0.07 (0.29)	1.13 (0.93)	1.25 (0.92)	1.28 (0.90)	0.04 (0.16)	0.02 (0.16)	0.03 (0.16)
Classroom-level quality									
CLASS mean		0.11 (0.16)	0.05 (0.15)		−0.06 (0.67)	−0.22 (0.65)		0.20* (0.10)	0.19† (0.10)
CLASS variability			−0.65*** (0.17)			−1.41 (0.89)			−0.13 (0.18)
Pseudo R^2	.63	.63	.64	.68	.68	.68	.52	.52	.53

Note. Unstandardized betas reported; standard errors in parentheses. CLASS = Classroom Assessment Scoring System; QRIS = Quality Rating and Improvement System; CDA = Child Development Associate.
†$p < .10$.*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 5

Results From Regression Models With Instructional Support Predicting School Readiness Outcomes

	Instructional Support								
	Literacy			Language			Math		
Variable	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Child-level controls									
Age in months	0.03 (0.04)	0.03 (0.04)	0.03 (0.04)	0.45** (0.13)	0.47*** (0.13)	0.46*** (0.13)	0.06* (0.03)	0.07* (0.03)	0.07* (0.03)
Female	0.19 (0.27)	0.20 (0.26)	0.21 (0.26)	0.34 (0.96)	0.45 (0.93)	0.47 (0.92)	−0.13 (0.22)	−0.13 (0.22)	−0.13 (0.22)
Cohort	−0.64** (0.23)	−0.72** (0.24)	−0.69** (0.26)	−0.33 (0.86)	−0.85 (0.91)	−0.61 (0.98)	−0.26 (0.19)	−0.30 (0.20)	−0.36† (0.21)
Fall autoregressor	0.93*** (0.06)	0.93*** (0.06)	0.93*** (0.06)	0.82*** (0.03)	0.81*** (0.03)	0.81*** (0.03)	0.63*** (0.03)	0.62*** (0.03)	0.63*** (0.03)
Classroom-level controls									
Teacher–child ratio	−4.35 (3.18)	−4.55 (3.20)	−4.26 (3.26)	−14.28 (10.68)	−15.25 (10.59)	−16.97 (10.79)	−3.43 (2.25)	−3.47 (2.29)	−2.69 (2.37)
Teacher education	0.21 (0.17)	0.20 (0.17)	0.10 (0.19)	0.71 (0.82)	0.78 (0.81)	1.02 (0.91)	−0.35* (0.15)	−0.32* (0.15)	−0.35* (0.17)
Teacher CDA credential	−0.71 (0.69)	−0.69 (0.68)	−0.72 (0.67)	−1.77 (2.97)	−1.71 (2.88)	−1.73 (2.84)	0.38 (0.42)	0.39 (0.43)	0.37 (0.43)
Teacher experience	0.03 (0.03)	0.03 (0.02)	0.02 (0.03)	−0.08 (0.19)	−0.11 (0.18)	−0.10 (0.18)	−0.01 (0.03)	−0.02 (0.03)	−0.02 (0.03)
QRIS level	−0.13 (0.29)	−0.08 (0.29)	−0.06 (0.29)	1.13 (0.93)	1.46 (0.89)	1.36 (0.91)	0.04 (0.16)	0.05 (0.16)	0.07 (0.16)
Classroom-level quality									
CLASS mean		−0.33† (0.19)	−0.29 (0.24)		−1.75** (0.59)	−1.85* (0.74)		−0.17 (0.13)	−0.07 (0.17)
CLASS variability			−0.12 (0.23)			0.03 (0.92)			−0.20 (0.25)
Pseudo R^2	.63	.64	.63	.68	.68	.68	.52	.52	.52

Note. Unstandardized betas reported; standard errors in parentheses. CLASS = Classroom Assessment Scoring System; QRIS = Quality Rating and Improvement System; CDA = Child Development Associate.
$^{\dagger}p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

are generally providing consistent, high-quality interactions throughout the day (Brock et al., 2018; Downer et al., 2010).

In line with our hypotheses, children had worse literacy performance at the end of the preschool year when they attended classrooms with more variable Classroom Organization, and better math performance when they attended classrooms with higher mean Classroom Organization. These findings are somewhat consistent with previous work that has documented a positive relation between mean Classroom Organization and math performance among elementary students (Sandilos et al., 2019), implying that children's math skills may be supported within classrooms where teachers are well prepared to facilitate learning activities and provide clear behavior expectations. Yet the exact conditions that are necessary to set the foundation for effective preschool math instruction should be further explored, given that this finding diminished in significance after including variability in the model. Notably, Sandilos et al. (2019) did not uncover significant links between variability in Classroom Organization and English Language Arts. The effect size of .09 that was observed in this study for variability in Classroom Organization and children's literacy is at the upper end of the published range in the existing literature (e.g., Perlman et al., 2016), suggesting that variability may be a more robust indicator of organizational quality for understanding development.

In general, results indicate that children's literacy skills may be negatively affected by teachers who are inconsistent in their behavior management strategies and deliver lessons with unclear instructions and minimal follow through. One explanation for this finding is that teaching strategies that promote literacy, such as directing children's attention to "code-based" aspects of oral and written language, may also require teachers to be intentional and consistent (e.g., deliver repeated dosage) to be effective (Hamre et al., 2010). Another possibility is that consistent and high-quality Classroom Organization may influence literacy through children's behavioral skills, such as their ability to pay attention, remember instructions, and follow rules (Downer et al., 2010). Future research should examine whether the nuances illustrated in the present study with regard to Classroom Organization exist in other data sets and explore the mechanisms that drive such effects, such as child engagement or withdrawal (Vitiello et al., 2012).

Finally, we did not find evidence of an association between variability in Instructional Support and any school readiness outcomes. We hypothesized that variability in this domain could be positively or negatively related to children's outcomes because of the competing theory and evidence (Brock et al., 2018; Sandilos et al., 2019). In this sample, the Instructional Support domain had the greatest amount of variability, and therefore, the strongest potential for explaining school readiness skills. Yet, only mean Instructional Support was negatively related to children's language skills. Previous research has mostly documented small positive effects of mean Instructional Support on school readiness, including vocabulary (Guo et al., 2010; Hamre et al., 2014; Hu et al., 2017; Hu et al., 2019; Mashburn et al., 2008). Our contradictory findings may be due to the low mean for Instructional Support, which in this sample, fell well below the threshold for what is considered high quality (Burchinal et al., 2010; Hatfield et al., 2016; Weiland et al., 2013). This suggests that most children were experiencing predominantly low-quality Instructional Support, with few instances of interactions in the mid-range. Alternatively, the unexpected result may indicate that teacher practices theorized to support language development within the Instructional Support domain, such as asking frequent questions, repeating and extending conversations, and using a variety of advanced words, may actually come at a cost to language development if teachers are not scaffolding at an appropriate or individualized level (Pentimonti et al., 2017). Indeed, there is some evidence that inexperienced children may not respond to intensive Instructional Support (Delaney & Krepps, 2021). In future work, it will be important to clarify what constitutes as significant changes in this domain in terms of promoting children's school readiness, particularly among diverse populations.

Overall, we uncovered little evidence that variability in CLASS scores and mean CLASS scores were consistently predictive of school readiness skills in the ways that we would expect and that would allow us to make many generalized conclusions. Our findings are largely at odds with the broader literature documenting the role of consistent (Brock & Curby, 2014; Curby et al., 2013) and high-quality (Araujo et al., 2016; Burchinal et al., 2008; Curby et al., 2009; Howes et al., 2008; Mashburn et al., 2008) teacher-child interactions for children's learning. However, they align with a few recent studies indicating small or null associations between mean CLASS scores and child outcomes (Guerrero-Rosada et al., 2021; Perlman et al., 2016; Weiland et al., 2013) and a single study demonstrating mostly nonsignificant relations between variability in CLASS scores and academic outcomes in elementary school (Sandilos et al., 2019). The general lack of significant findings across CLASS domains and school readiness outcomes could be an indication that our measure of variability is picking up on changes in lessons during the observation period that require teachers to fluctuate more in their interactions with children (Thorpe et al., 2020). Alternatively, the measure of variability may be capturing coder drift, which would likely not be indicative of classroom processes that matter for child outcomes (Burchinal, 2018). For instance, a recent G-study revealed that across all CLASS domains, less than 50% of the variability was attributable to the behaviors of teachers (Mantzicopoulus et al., 2018). Another explanation is that the CLASS domains are not reliably capturing their underlying dimensions (Gordon & Peng, 2020), thus inflating the

amount of variability in observer reports. Finally, it is possible that the CLASS, while practically informative, may not translate from the research context to complex classroom environments (Liu et al., 2019). In other words, the CLASS may be theoretically grounded but lack the ecological validity that is necessary to provide information about how children are learning from their interactions with their teachers—a hypothesis that is supported by emerging evidence that the CLASS does not always predict child outcomes (e.g., Guerrero-Rosada et al., 2021; Perlman et al., 2016; Weiland et al., 2013). Regardless, researchers and practitioners should consider whether the CLASS is an appropriate measure for their specific purpose and continue to explore how to capitalize on the information obtained from this popular tool.

Limitations, Strengths, and Future Directions

This study contributes to the early childhood literature by leveraging the strengths of a common measure of classroom quality (i.e., the CLASS) to investigate variability and its associations with children's school readiness skills. Despite the novel findings of this study, a few limitations along with their complementary strengths must be noted. First, it is possible that the indicators of variability are picking up on measurement error instead of meaningful variations in quality (Casabianca et al., 2015; Mashburn et al., 2014). In turn, inconsistencies within—and between—raters may mask true variability that is due to fluctuations in teaching practices. The CLASS has been criticized for its low reliability criteria (Burchinal, 2018). To illustrate, differences in raters have been found to account for 22% to 32% of the variance in Classroom Organization scores (Mantzicopoulos et al., 2018; Praetorius et al., 2014). If the findings from the present study are reflective of poor interrater reliability in observer ratings and are not representative of true variability, this points to a potential measurement issue of the CLASS training and protocols that should be revisited. Unfortunately, this study lacks access to data that could provide answers to these questions. Therefore, an important direction for future research is to explore whether individual observers impart their own biases as they rate single classrooms over time and understand why classroom quality seems to vary. This may be particularly critical to consider with respect to the Instructional Support domain because it is most challenging for raters to reliably score (Styck et al., 2021). Specifically, G-studies are useful for informing the rigor that would be necessary to achieve reliability (i.e., stability) in classroom quality ratings, such as the number of observation intervals, days, and raters (Mantzicopoulos et al., 2018; Mashburn et al., 2014). Given that mean scores on the CLASS vary from day to day (Buell et al., 2017), so too may variability in CLASS scores.

Another weakness of the CLASS is that it only captures one brief snapshot of quality across the many days and interactions that children experience. Researchers argue that a single measure of complex classroom processes may not be adequate (Weiland et al., 2013). Future work should attempt to replicate findings using more precise measures of teacher-child interactions, such as those that examine the quality of teacher-child interactions for individual children (Bohlmann et al., 2019; Downer et al., 2011). Another promising future direction may be exploring different time metrics of variability to investigate whether micro- versus macroinconsistencies in quality have greater significance for children's learning. At a microlevel, this may include rating the frequency of classroom strategies and interactions across multiple intervals and days (Kettler et al., 2019). At a macrolevel, it could involve charting change over the year or multiple years. For instance, gains in Instructional Support over the preschool year have been shown to be related to children's literacy and inhibitory control (Goble et al., 2019), and researchers have demonstrated that consistently high Instructional Support from preschool to kindergarten promotes language and literacy development (Cash et al., 2019). Moreover, examining multiple ways of measuring variability within a single day (e.g., Wang et al., 2020) and their consequences for children's learning is an important next step as different approaches may yield different conclusions and interpretations.

The CLASS is a global measure of quality, and while its versatility can be viewed as a strength, the lack of specificity also leaves room for subjectivity. For example, one dimension within the Instructional Support domain measures the quality of language modeling. However, the measure itself is not designed to assess the quality of instruction during literacy activities exclusively. Although preschool teachers spend a large amount of their day focusing on language and literacy (Early et al., 2010), it is possible that instructional quality may vary depending on the context and content of instruction (Rimm-Kaufman et al., 2005). Indeed, one study found that CLASS scores were influenced by the content of activities (Thorpe et al., 2020). Although our data restricted such analysis, researchers should carefully examine whether variability in quality is consistent across different types of activities in preschool and what the implications of variability in these various settings are for children's school readiness.

Last, while we view the use of a sample from families from low incomes as an asset because these children are most in need of high-quality instruction, it is important to acknowledge that our results may be specific to this population. Prior work has demonstrated that quality is generally lower in classrooms serving children from families with low incomes (Pianta et al., 2007). Therefore, future research should aim to replicate these associations in diverse preschool classrooms to make broader conclusions about the

generalizability of these findings. Moreover, the lack of access to classroom and teacher information inherent in the study design lends itself to the potential for omitted variable bias and Type I error. Although we took a conservative approach to interpreting the practical significance of findings, it will be necessary to examine these questions with a dataset that allows researchers to account for the many factors that may influence variability in teacher–child interactions, such as teacher stress, to better understand the mechanisms linking classroom quality to child outcomes (Li Grining et al., 2010).

Implications

Results from the present study have implications for research and practice. The CLASS is more frequently being incorporated within state QRIS (Sabol & Pianta, 2015). Such widespread use of the CLASS in research and practice has generated strong financial and political stakes (Tout et al., 2009), including informing policy decisions about satisfactory thresholds in Head Start and state-prekindergarten. Like others, we caution against using average scores or defaulting to guidelines set by thresholds on observational measures like the CLASS as a sole indicator of effective or ineffective teachers and for the purpose of providing merit (Good & Lavigne, 2015; Mashburn, 2017; Mantzicopoulos et al., 2018). Instead, we recommend that the nuances in teacher-child interactions also be considered when tracking progress across several measures that assess instructional effectiveness.

Our findings for Classroom Organization provide some preliminary evidence that efforts to reduce achievement gaps by improving quality, such as providing access to state-funded prekindergarten programs and administering QRIS, may also want to measure and monitor classrooms with high variability. Recent evidence suggests that it is possible to improve CLASS scores over time with targeted investments, and ECE programs who score in the mid-range on the CLASS may have the greatest potential for growth (Bassok et al., 2021). Thus, in addition to increasing the quality of teacher-child interactions to improve children's school readiness (e.g., Wasik & Hindman, 2011; Markowitz et al., 2018; Mashburn et al., 2015), continuous improvement initiatives should also consider supporting programs in sustaining the high-quality interactions they already show the capacity for. Focusing on creating a stable classroom environment in terms of structure, expectations, and management could be an essential approach to professional development for teachers that is commonly overlooked. Of course, more work is needed to understand whether and how these results hold up across samples before making any firm conclusions regarding their implications.

Finally, although we are encouraged by the fact that state and federal initiatives have started to incorporate theoretically and empirically informed practices into their policies, we share the concern of others about using global assessments to meet policy goals (Burchinal, 2018; Gordon & Peng, 2020). Specifically, widespread use the CLASS for these purposes has proven feasible; however, some precision in detecting meaningful effects has been compromised along the way (Pianta et al., 2020). This suggests the need to continue developing and refining measures of classroom quality that align with the evidence on what we know works for individual children and can be implemented at scale without losing integrity.

Conclusions

The CLASS is predicated on the assumption that consistent, high-quality interactions are essential components of classroom quality that shape children's development. Most of the models in the present study, however, yielded null findings or results that run contrary to conclusions drawn in the broader literature. The present study advances the field by illustrating that two conceptualizations of classroom quality previously shown to influence children's outcomes do not produce the same anticipated effects across all large samples, CLASS domains, and school readiness outcomes. Furthermore, findings suggest that variability may be a more robust indicator of quality in the domain of Classroom Organization than the mean when considering children's growth in literacy. Results have important implications for professional development and practice, particularly within the context of state-funded prekindergarten. Although the CLASS provides a theoretically grounded approach to defining classroom quality, there is still more work to be done to better understand and improve on the validity and practical significance of this widely administered measures of quality. An important first step is to build a more comprehensive body of knowledge around what information we can expect to obtain from global observations of classroom quality, derived from both mean scores and variability in scores. This will allow us to achieve a more realistic understanding of the circumstances under which children benefit within existing frameworks and enable the development of complementary measures of quality that fill the gaps.

Appendix

TABLE A1
Teacher Demographics

Variable	*n*	*M*	*SD*	Range
Days children attend per week	78	4.98	0.13	4–5
Hours children attend per day	78	7.97	1.75	1–12
Years teaching preschool	76	6.80	7.47	0–38
Teacher Education				
<Eighth grade	1	0.01	0.11	0–1
Some high school	10	0.13	0.33	0–1
High school diploma or GED	42	0.53	0.50	0–1
Trade school	23	0.29	0.46	0–1
Some college	4	0.05	0.22	0–1
Teacher has Child Development Associate	79	0.37	0.49	0–1
Teacher has teaching license from state	80	0.08	0.27	0–1

Note. GED = General Education Development.

Funding

This study was funded by Indiana's Family and Social Services Administration [contract # F1-79-15-PK-0374 and contract #00000000000000000000026332]. The opinions expressed here are those of the authors and do not represent view of the institution.

ORCID iDs

Jennifer K. Finders https://orcid.org/0000-0001-8368-4302
Robert J. Duncan https://orcid.org/0000-0001-6900-0322

References

Acock, A. C. (2012). What to do about missing values. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (pp. 27–50). American Psychological Association. https://doi.org/10.1037/13621-002

Ainsworth, M. S. (1989). Attachments beyond infancy. *American Psychologist*, *44*(4), 709–716. https://doi.org/10.1037/0003-066X.44.4.709

Ainsworth, M. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Lawrence Erlbaum.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics*, *131*(3), 1415–1453. https://doi.org/10.1093/qje/qjw016

Bailey, R. P., Hillman, C., Arent, S., & Petitpas, A. (2013). Physical activity: An underestimated investment in human capital. *Journal of Physical Activity and Health*, *10*(3), 289–308. https://doi.org/10.1123/jpah.10.3.289

Barnett, W. S., Jung, K., Friedman-Krauss, A., Frede, E. C., Nores, M., Hustedt, J. T., Howes, C., & Daniel-Echols, M. (2018). State prekindergarten effects on early learning at kindergarten entry: An analysis of eight state programs. *AERA Open*, *4*(2). https://doi.org/10.1177/2332858418766291

Bassok, D., Magouirk, P., & Markowitz, A. J. (2021). Systemwide quality improvement in early childhood education: Evidence from Louisiana. *AERA Open*, *7*. https://doi.org/10.1177/23328584211011610

Bohlmann, N. L., Downer, J. T., Williford, A. P., Maier, M. F., Booren, L. M., & Howes, C. (2019). Observing children's engagement: Examining factorial validity of the inCLASS across demographic groups. *Journal of Applied Developmental Psychology*, *60*, 166–176. https://doi.org/10.1016/j.appdev.2018.08.007

Bornstein, M. H., Haynes, M. O., & Painter, K. M. (1998). Sources of child vocabulary competence: A multivariate model. *Journal of Child Language*, *25*(2), 367–393. https://doi.org/10.1017/S0305000998003456

Brock, L. L., & Curby, T. W. (2014). Emotional support consistency and teacher–child relationships forecast social competence and problem behaviors in prekindergarten and kindergarten. *Early Education and Development*, *25*(5), 661–680. https://doi.org/10.1080/10409289.2014.866020

Brock, L. L., Curby, T. W., & Cannell-Cordier, A. L. (2018). Consistency in children's classroom experiences and implications for early childhood development. In A. J. Mashburn, J. LoCasale-Crouch, & K. C. Pears (Eds.), *Kindergarten transition and readiness* (pp. 59-83). Springer International.

Bronfenbrenner, U., & Morris, P. A. (1998). The ecology of developmental processes. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development* (pp. 993–1028). John Wiley.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner & W. Damon (Eds.), *Handbook of child psychology: Theoretical models of human development* (pp. 793–828). John Wiley.

Buell, M., Han, M., & Vukelich, C. (2017). Factors affecting variance in Classroom Assessment Scoring System scores: Season, context, and classroom composition. *Early Child Development and Care*, *187*(11), 1635–1648. https://doi.org/10.1080/03004430.2016.1178245

Burchinal, M. (2018). Measuring early care and education quality. *Child Development Perspectives*, *12*(1), 3–9. https://doi.org/10.1111/cdep.12260

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Development Science*, *12*(3), 140–153. https://doi.org/10.1080/10888690802199418

Burchinal, M., McCartney, K., Steinberg, L., Crosnoe, R., Friedman, S. L., McLoyd, V., & Pianta, R., & NICHD Early Child Care Research Network. (2011). Examining the Black–White achievement gap among low-income children using the NICHD study of early child care and youth development. *Child Development*, *82*(5), 1404–1420. https://doi.org/10.1111/j.1467-8624.2011.01620.x

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*, *25*(2), 166–176. https://doi.org/10.1016/j.ecresq.2009.10.004

Burchinal, M., Vernon-Feagans, L., Vitiello, V., & Greenberg, M., & Family Life Project Key Investigators. (2014). Thresholds in the association between child care quality and child outcomes in rural preschool children. *Early Childhood Research Quarterly*, *29*(1), 41–51. https://doi.org/10.1016/j.ecresq.2013.09.004

Cabell, S. Q., DeCoster, J., LoCasale-Crouch, J., Hamre, B. K., & Pianta, R. C. (2013). Variation in the effectiveness of instructional interactions across preschool classroom settings and learning activities. *Early Childhood Research Quarterly*, *28*(4), 820–830. https://doi.org/10.1016/j.ecresq.2013.07.007

Cameron, C. E., Connor, C. M., & Morrison, F. J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology*, *43*(1), 61–85. https://doi.org/10.1016/j.jsp.2004.12.002

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311–337. https://doi.org/10.1177/0013164414539163

Cash, A. H., Ansari, A., Grimm, K. J., & Pianta, R. C. (2019). Power of two: The impact of 2 years of high quality teacher child interactions. *Early Education and Development*, *30*(1), 60–81. https://doi.org/10.1080/10409289.2018.1535153

Cash, A. H., & Pianta, R. C. (2014). The role of scheduling in observing teacher–child interactions. *School Psychology Review*, *43*(4), 428–449. https://doi.org/10.1080/02796015.2014.12087414

Cohen, J., Ruzek, E., & Sandilos, L. (2018). Does teaching quality cross subjects? Exploring consistency in elementary teacher practice across subjects. *AERA Open*, *4*(3). https://doi.org/10.1177/2332858418794492

Curby, T. W., Brock, L. L., & Hamre, B. K. (2013). Teachers' emotional support consistency predicts children's achievement gains and social skills. *Early Education & Development*, *24*(3), 292–309. https://doi.org/10.1080/10409289.2012.665760

Curby, T. W., Grimm, K. J., & Pianta, R. C. (2010). Stability and change in early childhood classroom interactions during the first two hours of a day. *Early Childhood Research Quarterly*, *25*(3), 373–384. https://doi.org/10.1016/j.ecresq.2010.02.004

Curby, T. W., Rimm-Kaufman, S. E., & Cameron-Ponitz, C. C. (2009). Teacher–child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology*, *101*(4), 912–925. https://doi.org/10.1037/a0016647

Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., Hamre, B., & Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. *Elementary School Journal*, *112*(1), 16–37. https://doi.org/10.1086/660682

Davis, E. A., & Miyake, N. (2004). Explorations of scaffolding in complex classroom systems. *Journal of the Learning Sciences*, *13*, 265–272. https://doi.org/10.1207/s15327809jls1303_1

Delaney, K. K., & Krepps, K. (2021). Exploring Head Start teacher and leader perceptions of the Pre-K Classroom Assessment Scoring System as a part of the Head Start Designation Renewal System. *Early Childhood Research Quarterly*, *55*, 214–229. https://doi.org/10.1016/j.ecresq.2020.09.013

Downer, J. T., Booren, L. M., Hamre, B., Pianta, R. C., & Williford, A. (2011). *The Individualized Classroom Assessment Scoring (inCLASS)* [Database record]. Curry School of Education, University of Virginia, Charlottesville, VA. https://doi.org/10.1037/t76724-000

Downer, J. T., Sabol, T. J., & Hamre, B. (2010). Teacher–child interactions in the classroom: Toward a theory of within and cross-domain links to children's developmental outcomes. *Early Education & Development*, *21*(5), 699–723. https://doi.org/10.1080/10409289.2010

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, *50*(11), 2417. https://doi.org/10.1037/a0037996

Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test* (4th ed.). Pearson Education.

Durham, R. E., Farkas, G., Hammer, C. S., Tomblin, J. B., & Catts, H. W. (2007). Kindergarten oral language skill: A key variable in the intergenerational transmission of socioeconomic status. *Research in Social Stratification and Mobility*, *25*(4), 294–305. https://doi.org/10.1016/j.rssm.2007.03.001

Early, D. M., Iruka, I. U., Ritchie, S., Barbarin, O. A., Winn, D. M. C., Crawford, G. M., Frome, P. M., Clifford, R. M., Burchinal, M., Howes, C., Bryant, D. M., & Pianta, R. C. (2010). How do pre-kindergarteners spend their time? Gender, ethnicity, and income as predictors of experiences in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *25*(2), 177–193. https://doi.org/10.1016/j.ecresq.2009.10.003

Early, D. M., Maxwell, K. L., Ponder, B. D., & Pan, Y. (2017). Improving teacher-child interactions: A randomized controlled trial of Making the Most of Classroom Interactions and My Teaching Partner professional development models. *Early Childhood Research Quarterly*, *38*, 57–70.

Elicker, J. G., Langill, C. C., Ruprecht, K. M., Lewsader, J., & Anderson, T. (2011). *Paths to QUALITY, Indiana's Child Care Quality Rating and Improvement System: Final Report* (Technical Report 3). Purdue University. https://doi.org/10.13140/2.1.2488.7044

Friedman-Krauss, A. H., Barnett, W. S., Garver, K.A., Hodges, K.S., Weisenfeld, G. G., & Gardiner, B.A. (2020). *The State of*

Preschool 2019: State Preschool Yearbook. National Institute for Early Education Research.

Goble, P., Sandilos, L. E., & Pianta, R. C. (2019). Gains in teacher-child interaction quality and children's school readiness skills: Does it matter where teachers start? *Journal of School Psychology*, *73*, 101–113. https://doi.org/10.1016/j.jsp.2019.03.006

Good, T. L., & Lavigne, A. L. (2015). Rating teachers cheaper, faster, and better: Not so fast. *Journal of Teacher Education*, *66*(3), 288–293. https://doi.org/10.1177/0022487115574292

Gordon, R. A., & Peng, F. (2020). Evidence regarding the domains of the CLASS PreK in Head Start classrooms. *Early Childhood Research Quarterly*, *53*, 23–39. https://doi.org/10.1016/j.ecresq.2020.01.008

Guerrero-Rosada, P., Weiland, C., McCormick, M., Hsueh, J., Sachs, J., Snow, C., & Maier, M. (2021). Null relations between CLASS scores and gains in children's language, math, and executive function skills: A replication and extension study. *Early Childhood Research Quarterly*, *54*, 1–12. https://doi.org/10.1016/j.ecresq.2020.07.009

Guo, Y., Piasta, S. B., Justice, L. M., & Kaderavek, J. N. (2010). Relations among preschool teachers' self-efficacy, classroom quality, and children's language and literacy gains. *Teaching and Teacher Education*, *26*(4), 1094–1103. https://doi.org/10.1016/j.tate.2009.11.005

Hamre, B. K. (2014). Teachers' daily interactions with children: An essential ingredient in effective early childhood programs. *Child Development Perspectives*, *8*(4), 223–230.

Hamre, B. K., 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. https://doi.org/10.1111/cdev.12184

Hamre, B. K., Justice, L. M., Pianta, R. C., Kilday, C., Sweeney, B., Downer, J. T., & Leach, A. (2010). Implementation fidelity of MyTeachingPartner literacy and language activities: Association with preschoolers' language and literacy growth. *Early Childhood Research Quarterly*, *25*(3), 329–347. https://doi.org/10.1016/j.ecresq.2009.07.002

Hamre, B. K., & Pianta, R. C. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, *72*(2), 625–638. https://doi.org/10.1111/1467-8624.00301

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. L. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 49–83). Paul H Brookes.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale*. Teachers College Press.

Hatfield, B. E., Burchinal, M. R., Pianta, R. C., & Sideris, J. (2016). Thresholds in the association between quality of teacher–child interactions and preschool children's school readiness skills. *Early Childhood Research Quarterly*, *36*, 561–571. https://doi.org/10.1016/j.ecresq.2015.09.005

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64. https://doi.org/10.3102/0013189X12437203

Hong, S. L. S., Sabol, T. J., Burchinal, M. R., Tarullo, L., Zaslow, M., & Peisner-Feinberg, E. S. (2019). ECE quality indicators and child outcomes: Analyses of six large child care studies. *Early Childhood Research Quarterly*, *49*, 202–217. https://doi.org/10.1016/j.ecresq.2019.06.009

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, *23*(1), 27–50. https://doi.org/10.1016/j.ecresq.2007.05.002

Hu, B. Y., Fan, X., Wu, Y., LoCasale-Crouch, J., & Song, Z. (2019). Contributions of teacher–child interaction quality to Chinese children's development in the early childhood years. *Early Education and Development*, *30*(2), 159–177. https://doi.org/10.1080/10409289.2018.1544809

Hu, B. Y., Fan, X., Wu, Z., LoCasale-Crouch, J., Yang, N., & Zhang, J. (2017). Teacher-child interactions and children's cognitive and social skills in Chinese preschool classrooms. *Children and Youth Services Review*, *79*, 78–86. https://doi.org/10.1016/j.childyouth.2017.05.028

Kettler, R. J., Reddy, L. A., Glover, T. A., & Kurz, A. (2019). Bridging the gap: Classroom Strategies Assessment System–Observer Form. *Assessment for Effective Intervention*, *44*(2), 120–122. https://doi.org/10.1177/1534508417747391

Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., Ruzek, E. A., & Howes, C. (2013). Preschool center quality and school readiness: Quality effects and variation by demographic and child characteristics. *Child Development*, *84*(4), 1171–1190. https://doi.org/10.1111/cdev.12048

Kook, J. F., & Greenfield, D. B. (2020). Examining variation in the quality of instructional interaction across teacher-directed activities in head start classrooms. *Journal of Early Childhood Research*, *19*(2), 128–144. https://doi.org/10.1177/1476718X20942956

La Paro, K. M., Hamre, B. K., Locasale-Crouch, J., Pianta, R. C., Bryant, D., Early, D., Clifford, R., Barbarin, O., Howes, C., & Burchinal, M. (2009). Quality in kindergarten classrooms: Observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Education and Development*, *20*(4), 657–692.

La Paro, K. M., Pianta, R., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *Elementary School Journal*, *104*(5), 409–426. https://doi.org/10.1086/499760

Lahti, M., Elicker, J., Zellman, G., & Fiene, R. (2015). Approaches to validating child care quality rating and improvement systems (QRIS): Results from two states with similar QRIS type designs. *Early Childhood Research Quarterly*, *30*(B), 280–290. https://doi.org/10.1016/j.ecresq.2014.04.005

Leyva, D., Weiland, C., Barata, M., Yoshikawa, H., Snow, C., Treviño, E., & Rolla, A. (2015). Teacher–child interactions in Chile and their associations with prekindergarten outcomes. *Child Development*, *86*(3), 781–799. https://doi.org/10.1111/cdev.12342

Li Grining, C., Raver, C. C., Champion, K., Sardin, L., Metzger, M., & Jones, S. M. (2010). Understanding and improving classroom emotional climate and behavior management in the "real world": The role of Head Start teachers' psychosocial stressors. *Early Education and Development*, *21*(1), 65–94. https://doi.org/10.1080/10409280902783509

Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, *31*(1), 61–95.

LoCasale-Crouch, J., Jamil, F., Pianta, R. C., Rudasill, K. M., & DeCoster, J. (2018). Observed quality and consistency of fifth graders' teacher–student interactions: Associations with feelings, engagement, and performance in school. *SAGE Open*, *8*(3). https://doi.org/10.1177/2158244018794774.

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: a generalizability study comparing the framework for teaching and the classroom assessment scoring system. *Educational Assessment*, *23*(1), 24–46. https://doi.org/10.1080/10627197.2017.1408407

Markowitz, A. J., Bassok, D., & Hamre, B. (2018). Leveraging developmental insights to improve early childhood education. *Child Development Perspectives*, *12*(2), 87–92. https://doi.org/10.1111/cdep.12266

Mashburn, A. J. (2017). Evaluating the validity of classroom observations in the Head Start Designation Renewal System. *Educational Psychologist*, *52*(1), 38–49. https://doi.org/10.1080/00461520.2016.1207539

Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science*, *15*(2), 146–155. https://doi.org/10.1007/s11121-012-0357-3

Mashburn, A. J., Justice, L., Mcginty, A., & Slocum, L. (2015). The impacts of a scalable intervention on the language and literacy development of rural pre-kindergartners. *Applied Developmental Science*, *20*(1), 61–78. https://doi.org/10.1080/10888691.2015.1051622

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*(3), 732–749. https://doi.org/10.1111/j.1467-8624.2008.01154.x

Patrick, H., & Mantzicopoulos, P. (2016). Is effective teaching stable? *Journal of Experimental Education*, *84*(1), 23–47.

Pentimonti, J. M., Justice, L. M., Yeomans-Maldonado, G., McGinty, A. S., Slocum, L., & O'Connell, A. (2017). Teachers' use of high-and low-support scaffolding strategies to differentiate language instruction in high-risk/economically disadvantaged settings. *Journal of Early Intervention*, *39*(2), 125–146. https://doi.org/10.1177/1053815117700865

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*, Article e0167660. https://doi.org/10.1371/journal.pone.0167660

Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. (2007). Opportunities to learn in America's elementary classrooms. *Science*, *315*(5820), 1795–1796. https://doi.org/10.1126/science.1139719

Pianta, R. C., Hamre, B. K., & Nguyen, T. (2020). Measuring and improving quality in early care and education. *Early Childhood Research Quarterly*, *51*, 285–287. https://doi.org/10.1016/j.ecresq.2019.10.013

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual Pre-K*. Paul H Brookes.

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2-12.

Qi, C. H., Zieher, A., Lee Van Horn, M., Bulotsky-Shearer, R., & Carta, J. (2019). Language skills, behaviour problems, and classroom emotional support among preschool children from low-income families. *Early Child Development and Care*, *190*(14), 2278–2290.

Rimm-Kaufman, S. E., La Paro, K. M., Downer, J. T., & Pianta, R. C. (2005). The contribution of classroom setting and quality of instruction to children's behavior in kindergarten classrooms. *Elementary School Journal*, *105*(4), 377–394.

Sabol, T., & Pianta, R. (2015). Validating Virginia's quality rating and improvement system among state-funded pre-kindergarten programs. *Early Childhood Research Quarterly*, *30*(B), 183–198. https://doi.org/10.1016/j.ecresq.2014.03.004

Sandilos, L. E., Sims, W. A., Norwalk, K. E., & Reddy, L. A. (2019). Converging on quality: Examining multiple measures of teaching effectiveness. *Journal of School Psychology*, *74*, 10–28. https://doi.org/10.1016/j.jsp.2019.05.004

Schrank, F. A., McGrew, K. S., Mather, N., Wendling, B. J., & LaForte, E. M. (2014). *Woodcock-Johnson IV tests of achievement: Woodcock-Johnson IV tests of cognitive abilities*. Riverside.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, *46*(4), 553–611. https://doi.org/10.3102/00346543046004553

Snow, C. E., & Matthews, T. J. (2016). Reading and language in the early grades. *Future of Children*, *26*(2), 57–74. https://doi.org/10.1353/foc.2016.0012

Song, S., Su, M., Kang, C., Liu, H., Zhang, Y., McBride-Chang, C., Tardif, T., Li, H., Liang, W., Zhang, Z., & Shu, H. (2015). Tracing children's vocabulary development from preschool through the school-age years: An 8-year longitudinal study. *Developmental Science*, *18*(1), 119–131. https://doi.org/10.1111/desc.12190

Stapleton, L. M., McNeish, D. M., & Yang, J. S. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist, 51*(3–4), 317–330.

StataCorp. (2015). *Stata Statistical Software: Release 14*.

Styck, K. M., Anthony, C. J., Sandilos, L. E., & DiPerna, J. C. (2021). Examining rater effects on the Classroom Assessment Scoring System. *Child Development*, *92*(3), 976–993. https://doi.org/10.1111/cdev.13460

Thorpe, K., Rankin, P., Beatton, T., Houen, S., Sandi, M., Siraj, I., & Staton, S. (2020). The when and what of measuring ECE quality: Analysis of variation in the Classroom Assessment Scoring System (CLASS) across the ECE day. *Early Childhood Research Quarterly*, *53*, 274–286.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009, May). *Issues for the Next Decade of Quality Rating and Improvement Systems* (Issue Brief 3). Child Trends and the U.S. Department

of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. https://www.childtrends.org/wp-content/uploads/2009/05/Child_Trends-2009_5_19_RB_QualityRating.pdf

Tran, H., & Winsler, A. (2011). Teacher and center stability and school readiness among low-income, ethnically diverse children in subsidized, center-based child care. *Children and Youth Services Review*, *33*(11), 2241–2252. https://doi.org/10.1016/j.childyouth.2011.07.008

Turnbull, K. P., Anthony, A. B., Justice, L., & Bowles, R. (2009). Preschoolers' exposure to language stimulation in classrooms serving at-risk children: The contribution of group size and activity context. *Early Education and Development*, *20*(1), 53–79.

U.S. Department of Health and Human Services, Administration for Children & Families, Office of Head Start. (2020). *Head Start Program Performance Standards*. https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/hspps-final.pdf

Villarreal, V. (2015). Test Review: Woodcock-Johnson IV Tests of Achievement. *Journal of Psychoeducational Assessment*, *33*(4), 391–398. https://doi.org/10.1177/0734282915569447

Vitiello, V. E., Booren, L. M., Downer, J. T., & Williford, A. P. (2012). Variation in children's classroom engagement throughout a day in preschool: Relations to classroom and child factors. *Early Childhood Research Quarterly*, *27*(2), 210–220. https://doi.org/10.1016/j.ecresq.2011.08.005

von Hippel, P., Workman, J., & Downey, D. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "Are Schools the Great Equalizer?" *Sociology of Education*, *91*(4), 323–357. https://doi.org/10.1177/0038040718801760

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*(4), 509–519.

Vygotsky, L. S. (1991). Genesis of the higher mental functions. In P. Light, S. Sheldon, & M. Woodhad (Eds.), *Learning to think* (pp. 32–41). Taylor & Francis/Routledge.

Wang, S., Hu, B. Y., Curby, T., & Fan, X. (2020). Multiple approaches for assessing within-day stability in teacher-child interactions. *Early Education and Development*, *32*(4), 553–571.

Wasik, B. A., & Hindman, A. H. (2011). Improving vocabulary and pre-literacy skills of at-risk preschoolers through teacher professional development. *Journal of Educational Psychology*, *103*(2), 455–469. https://doi.org/10.1037/a0023067

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*(2), 199–209. https://doi.org/10.1016/j.ecresq.2012.12.002

Zinsser, K. M., Bailey, C. S., Curby, T. W., Denham, S. A., & Bassett, H. H. (2013). Exploring the predictable classroom: Preschool teacher stress, emotional supportiveness, and students' social-emotional behavior in private and Head Start classrooms. *NHSA Dialog*, *16*(2).

Authors

JENNIFER K. FINDERS is an assistant professor at Purdue University in Human Development and Family Studies. Her research focuses on the role of early care and education programs in promoting cognitive, behavioral, and social and emotional learning among children from diverse backgrounds.

ADASSA BUDREVICH is a senior evaluator for Washington State Department of Children, Youth, and Families. Her research focuses on evaluating how early care and education programs, workforce development, and classroom quality optimize outcomes for children furthest from opportunity.

ROBERT J. DUNCAN is an assistant professor at Purdue University in Human Development and Family Studies and Public Health. His research focuses on understanding children's development of academic skills, executive function, social-emotional skills, and fine/gross motor skills.

DAVID J. PURPURA is an associate professor at Purdue University in Human Development and Family Studies. His research focuses on understanding how young children in preschool through third grade learn math and how to identify children at-risk for later math difficulties.

JAMES ELICKER is professor emeritus at Purdue University in Human Development and Family Studies. His research has focused on evaluating quality improvement efforts in early care and education.

SARA A. SCHMITT is an associate professor at Purdue University in Human Development and Family Studies. Her research focuses on examining the individual and contextual factors that contribute to growth in self-regulation (e.g., executive function) and school readiness (e.g., social-emotional competence, early academic skills).