ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 24-897

The Ups and Downs of Classroom Quality Over the Preschool Year and Relations to Children's School Readiness

Kathryn E. Gonzalez
Mathematica

Olivia Healy
Elon University

Luke Miratrix
Harvard University

Terri J. Sabol
Northwestern University

Despite considerable evidence on the links between average classroom quality and children's learning, the importance of variation in quality is not well understood. We examined whether three measures of variation in observed classroom quality over the school year – overall variation in quality, teacher-specific trends in quality, and instability in quality – were associated with children's language, literacy, and regulatory outcomes. We also examined whether variation in quality was associated with teachers' participation in coaching. Overall variation and instability in emotional support and classroom organization over the year were negatively associated with children's regulatory and literacy outcomes. Participation in coaching was linked to increased variation only in instructional support. We discuss implications for policies focused on improving classroom quality.

VERSION: January 2024

Suggested citation: Gonzalez, Kathryn E., Olivia Healy, Luke Miratrix, and Terri J. Sabol. (2024). The Ups and Downs of Classroom Quality Over the Preschool Year and Relations to Children's School Readiness. (EdWorkingPaper: 24-897). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/stac-vs74

The Ups and Downs of Classroom Quality Over the Preschool Year and Relations to Children's School Readiness

Kathryn E. Gonzalez
Mathematica

Olivia Healy
Department of Economics, Elon University

Luke Miratrix
Harvard Graduate School of Education

Terri J. Sabol
School of Education and Social Policy, Northwestern University

Abstract

Despite considerable evidence on the links between average classroom quality and children's learning, the importance of *variation* in quality is not well understood. We examined whether three measures of variation in observed classroom quality over the school year – overall variation in quality, *teacher-specific trends* in quality, and *instability* in quality – were associated with children's language, literacy, and regulatory outcomes. We also examined whether variation in quality was associated with teachers' participation in coaching. Overall variation and instability in emotional support and classroom organization over the year were negatively associated with children's regulatory and literacy outcomes. Participation in coaching was linked to increased variation only in instructional support. We discuss implications for policies focused on improving classroom quality.

Keywords: classroom quality; preschool; professional development

Research has documented the critical role of high-quality early learning environments for children's development (Burchinal, 2018; Mashburn et al., 2008). There is also growing recognition, drawn from developmental theory, that the quality of activities and interactions in children's classrooms – referred to as *process quality* – shapes children's learning (Hamre, Hatfield, Jamil, & Pianta, 2014). Early studies found that these aspects of quality, particularly the quality of interactions between teachers and children, were associated with child academic and social-emotional outcomes (Howes et al., 2008; Mashburn et al., 2008). One of the most popular classroom quality measures is the Classroom Assessment Scoring System (CLASS™; Pianta, La Paro & Hamre, 2008). The CLASS assesses the quality of interactions between teachers and children in three crucial domains, including emotional support, classroom organization, and instructional support. This measure shows promise for quantifying the elements of early education environments that matter for children's development, albeit with small to moderate effect sizes (Keys et al., 2013; Perlman et al., 2016; Vandenbroucke et al., 2018).

Encouraging evidence about the role of process quality for children's development led to the proliferation of observation-based accountability systems in early childhood education (ECE). As part of the federal Head Start Designation Renewal System (DRS), all Head Start programs are evaluated with the CLASS. Head Start grantees do not meet minimum scores on the CLASS domains risk losing funding (Office of Head Start, 2016). Similarly, state Quality Rating and Improve Systems (QRIS) have expanded as tools to standardize and improve ECE quality (Isner et al., 2011). QRIS provide quality ratings to participating programs, and most states incorporate observational quality measures (often the CLASS or the Early Childhood Environmental Rating Scale [ECERS; Harms, Clifford & Cryer, 1998]) in these ratings (QRIS Compendium, n.d.).

This turn towards observation-based accountability has been accompanied by widespread use of professional development – particularly coaching – focused on observational quality measures. For example, a web-mediated coaching program developed based on the CLASS framework has been widely used as a tool to improve the quality of teacher-child interactions (Early, Maxwell, Ponder, & Pan, 2017; Pianta, Mashburn, Downer, Hamre, & Justice, 2008). At the federal level, one study found that approximately three-quarters of Head Start grantees reported that their teachers participated in CLASS-focused coaching (Derrick-Mills et al., 2016). Similarly, many states offer trainings focused on observational quality measures such as the CLASS to programs in their QRIS (Isner et al., 2011; QRIS Compendium, n.d.).

The use of the CLASS in high stakes accountability systems has led to questions and research about the underlying assumptions around the way the tool is used in these systems (Mashburn, 2017). For instance, researchers have focused on the extent to which CLASS scores from one observer generalize to the entire set of observers (Mashburn et al., 2014), whether there is sufficient variation in the items or domains of the tool (Burchinal, 2018), and whether there might be threshold effects (Burchinal, Vandergrift, Pianta, & Mashburn, 2010). Other studies have raised questions about whether CLASS scores from the subset of classrooms generalize to all classrooms within schools (Burchinal, 2018; Sabol, Ross & Frost, 2020).

One major assumption that has not been explored is whether classroom quality is constant over the year, given that accountability systems typically only assess classrooms at one given time in the school year. Research from the parenting literature and family science shows that inconsistent parenting and that unstable home environments can inhibit children's development (Evans, 2006; Kohen, Leventhal, Dahinten, & McIntosh, 2000; Maccoby, 2000). One might ask whether instability in the quality of children's relationships with their teachers in

ECE settings may similarly have a negative effect on children. For example, children may not be able to build positive, supportive relationships with teachers who provide an emotionally supportive classroom environment on one day, but not on the next. Similarly, children may not be able to easily engage with learning content when classroom routines are unpredictable over the course of the year. Trends in classroom quality over time may also have implications for children. For example, declines in emotional support or classroom organization may indicate rising teacher stress. Alternatively, children whose cognitive skills are improving over the year may benefit from being in classrooms where the level of instruction is similarly increasing. Yet, to date, researchers have not been able to directly examine this question. Studies have typically measured quality based on observations conducted in a single day (Howes et al., 2008; Mashburn et al., 2008), which do not capture if quality fluctuates over time. Other studies have created aggregate quality measures based on observations spread across multiple days or weeks (Hamre et al., 2012; Mashburn et al., 2010), which can obscure potentially uneven experiences that young children may have in ECE classrooms. Much more research is needed to understand the effect of classroom variability for children's development. In this work we explore this question to inform our understanding of how ECE environments shape children's development, and to provide insight into how to create fair and accurate ECE accountability systems.

In the current paper, we test how variation in classroom quality over an academic year related to children's early learning outcomes. We rely on multiple observations of teacher-child interaction quality over the course of the preschool year to examine whether across-year variability in classroom quality predicted children's inhibitory control, self-regulation, language, and literacy outcomes. First, we test whether the amount of *overall variation* in classroom quality across the school year predict children's early learning outcomes. Second, we examine

whether children's outcomes are predicted by two distinct types of quality variation: variation due to teachers' trajectories of growth in quality across the year, which we refer to as *teacher-specific trends* in classroom quality, and variation due to fluctuations in quality around these trends, which we refer to as *instability* in classroom quality.

At the same time, coaching programs that provide personalized supports and target the quality of teacher-child interactions have been shown to increase average levels of process quality (Early et al., 2017; Pianta et al., 2008). However, it is not well-understood whether these programs create or minimize variation in quality in targeted domains. In our final research aim, we examine whether participation in web-based coaching was associated with distinct patterns of classroom quality over the year, including overall variation, trends, and instability in quality. The broad goal of our work is to both advance the science behind how variation in classroom quality may relate to children's learning, as well as inform questions about how to measure quality in ECE, including when and how frequently observations should be conducted to provide a complete picture of setting quality. We conclude with a discussion of the implications of our findings for research, policy, and practice.

Classroom Quality and Child Outcomes in Early Childhood Education

In ECE contexts, quality inputs fall under two broad domains. *Structural quality* refers to the resources and organization of resources in ECE settings; these features create the conditions for effective interactions and processes to support children's development. These interactions and processes, referred to as *process quality*, reflect the proximal, dynamic aspects of children's interactive activities and experiences in ECE settings (Cassidy et al., 2005). The theoretical importance of process quality is based in foundational frameworks that emphasize the role of bidirectional interactions between children and teachers (Bronfenbrenner, 1979; Bronfenbrenner

& Morris, 1998). Accordingly, widely-used measures of process quality capture whether there are sensitive, responsive interactions between teachers and children (Sabol & Pianta, 2012).

Early studies found that measures of process quality had more positive associations with child outcomes as compared to structural quality features (e.g., teacher qualifications and ratios; Early et al., 2007; Howes et al., 2008; Mashburn et al., 2008). With regards to the CLASS, emotional support reflects teachers' abilities to create positive classroom environments, and has been linked to children's social and emotional outcomes (Mashburn et al., 2008; Pianta et al., 2008). Classroom organization, which includes teachers' use of classroom routines and procedures to manage children's behavior, and instructional support, which reflects the use of strategies to promote children's cognitive and language development, have been linked to children's academic outcomes, self-regulation, and classroom engagement (Howes et al., 2008; Mashburn et al., 2008; Rimm-Kaufman et al., 2009).

Yet, recent research has yielded mixed findings regarding links between CLASS scores and child outcomes. Recent meta-analyses found only a few positive, significant associations between CLASS scores and child outcomes (Keys et al., 2013; Perlman et al., 2016; Vandenbroucke et al., 2018). In response to these mixed results, research has explored whether associations between CLASS scores and child outcomes depend on whether quality is above meaningful thresholds (Burchinal et al., 2010). Other research has hypothesized that, due to quality regulations in many ECE programs, there may be insufficient variation in widely-used quality measures to identify their associations with child outcomes (Weiland, Ulvestad, Sachs & Yoshikawa, 2013).

In the present study, we consider whether *variation* in classroom quality may explain these muted associations between observed quality and child outcomes. If there is considerable

variation in quality within classrooms from one observation to the next, measures of quality from one time point are likely to be noisy measures of overall classroom quality and show limited associations with child outcomes (Mashburn, 2017). In part due to this concern, other research has focused on quality measures aggregated across multiple observations conducted over the course of the year, which are thought to be more reliable estimates of teacher practice (Hamre et al., 2012; Mashburn et al., 2010). Yet, the degree of fluctuation in the quality of children's experiences may differ from one classroom to the next, even if those classrooms are characterized by the same average levels of quality. These fluctuations may themselves have implications for children's development.

Sources of Variation in Classroom Quality

Research points to two sources of variation in classroom quality that may be meaningful for children's development. First, teachers may show positive (or negative) growth in quality across the school year. This source of variation, which is referred to as *teacher-specific trends in CLASS scores,* is illustrated in Figure 1. Figure 1 presents CLASS scores for two teachers during the school year (selected from present analytic sample) who have approximately the same average level of quality throughout the year. However, one teacher displays a generally positive trajectory of change in CLASS scores and the other a negative trajectory.

After accounting for teachers' overall trajectories of growth, a second source of across-year quality variation comes from *instability in classroom quality* around teacher-specific trajectories of growth. For example, some classrooms may display large changes in quality from one day or week to the next; in other classrooms, deviations from the general trajectory of growth may be more stable. This second source of variation is illustrated in Figure 2, which presents a series of CLASS scores over time for two teachers (again selected from the present

analytic sample) with similar overall trajectories of growth. However, one teacher shows considerable variation in quality from observation to observation, while the second teacher's CLASS scores are grouped more tightly around her own growth curve.

Limited empirical evidence exists regarding the degree to which these types of changes in quality occur in ECE classrooms throughout the school year. In a study of novice preschool teachers, Buell and colleagues (2017) found some evidence of overall trajectories of growth: teachers' CLASS scores increased from fall to spring across multiple years of instruction. Similarly, Meyer and colleagues (2011) found that CLASS scores varied over the school year, but did not examine whether there were systematic increases or decreases in quality. More detailed evidence regarding how quality rises and falls over the year comes primarily from K-12 contexts. In a study of secondary school teachers, Malmberg and colleagues (2010) found that teachers' emotional support increased then declined over the school year, and that teachers' classroom organization increased linearly over time. Other research in secondary school contexts found declining levels of emotional support and instructional support across the school years (Casabianaca et al., 2013; Casabianca et al., 2015). Although we note that there are substantial differences between ECE and K-12 classroom contexts, and that the aspects of quality captured by the CLASS differ across these contexts, research in the K-12 context provides evidence to suggest that there may also be systematic trends in quality over the year in ECE settings.

Beyond time trends in classroom quality, research has found that quality varies considerably over the course of a single school day in ECE settings (e.g., Brock & Curby, 2014; Curby, Brock & Hamre, 2013). Other research has documented declines in instructional support and emotional support over the day (von Suchodoletz et al., 2014). Research conducted in K-12 classrooms using measures other than the CLASS has also found evidence of variation in quality

across lessons and occasions (e.g., Praetorius et al., 2014), suggesting that quality instability is not just an artifact of the CLASS.

Implications of Variation in Classroom for Child Outcomes

There is theoretical justification, and some empirical evidence, that trends and instability in classroom quality have implications for children's development. For example, Malmberg et al. (2010) suggest that patterns of positive growth over the year may be due to teachers' experiencing a "mastery effect" characterized by improvements in quality over the year. In contrast, negative growth over the year may be due to teachers experiencing a negative "reality shock" that leads to high initial levels of quality that then decline (Malmberg et al., 2010). Declines in classroom quality over the year due to rising teacher stress levels (or improvements due to teacher mastery) may have negative (or positive) consequences for children.

Existing research also suggests that relative stability in classroom quality may better support children's development as compared with classrooms with fluctuating levels of quality. For example, teachers' emotional support is theorized to promote children's social-emotional development (Hamre et al., 2014). Stability in emotional support across the year may be especially important for children's self-regulatory skills based on decades of evidence from the parenting literature that children thrive in predictable, regular environments (Evans, 2006; Martin, Razza, & Brooks-Gunn, 2011; Kohen, Leventhal, Dahinten, & McIntosh, 2000; Maccoby, 2000). We hypothesize that stable, supportive relationships between teachers and children may also be beneficial.

Limited empirical evidence shows short-term stability in quality (as measured by the CLASS) supports children's learning. Studies have found that within-day stability in classroom emotional support, based on observations of preschool classrooms during one or two days of the

school year, was positively associated with children's social and behavioral outcomes (Brock & Curby, 2014), and with gains in children's academic skills (Curby et al., 2013). However, there is little empirical evidence regarding whether trends in quality or instability in quality over the school year – including fluctuations in quality from day to day or week to week – relate to children's early learning outcomes.

Teacher Coaching and Variation in Classroom Quality

Coaching for in-service teachers is one of the most widely used tools by researchers and policymakers to improve classroom quality in ECE (Egert, Fukkink & Lont, 2018). Studies have generally found that teachers' participation in coaching can improve teacher practice across a range of measured domains (Egert et al., 2018). In terms of impacts on teacher-child interactions, Pianta and colleagues (2008) found that a web-based coaching program, *MyTeachingPartner*, improved the quality of teacher-child interactions across all three CLASS domains. This is consistent with other studies that showed coaching programs based on the CLASS framework can improve classroom quality (Early et al., 2017; Pianta et al., 2017).

However, little research to date has examined whether these programs generate variation in classroom quality, beyond impacts on the average quality. On the one hand, it is plausible that participation in coaching could help reduce variability in classroom quality. Professional development programs such as coaching may reduce teacher stress, thereby enabling teachers to provide a consistently high-quality classroom environment (Sandilos et al., 2018). The skills and knowledge teachers gain by participating in coaching programs may also prepare them to better respond to events that might otherwise disrupt the classroom environment. Alternatively, it is possible that teachers' participation in coaching could increase variation in quality. A central aim of coaching programs is to allow teachers to learn and implement new teaching strategies.

(Darling-Hammond, Hyler, & Gardner, 2017). Teachers participating in coach may be more likely to try and err as they learn to effectively implement new instructional approaches.

The impacts of professional development on variation in quality may also depend on teachers' initial levels of instructional effectiveness. Teachers with higher initial classroom quality may implement new practices more effectively. In this case, coaching could raise quality without raising instability. In contrast, teachers with less experience or lower classroom quality may have greater difficulty implementing new practices, leading to increased instability in the short term. Given the common use of coaching in ECE, understanding how coaching programs affect variation in quality, and how this variation relates to child outcomes, will provide new insight into possible added benefits or unintended drawbacks of current coaching programs.

The Present Study

In the present study, we examine the implications of variability in classroom quality over the preschool year for children's early development. Specifically, we examine whether children's inhibitory control, self-regulation, language, and literacy outcomes were predicted by (a) overall variation in classroom quality, (b) variation in classroom quality due to teacher-specific trends in quality, and (c) variation in classroom quality due to instability in quality, after controlling for average levels quality. We then examine whether variation in classroom quality, including overall variation, teacher-specific trends in quality, and instability in quality, differs based on whether teachers had been assigned to a coaching intervention or control condition.

Method

Data and Sample

Our data came from the National Center for Research on Early Childhood Education Professional Development Study (NCRECE PDS; Pianta & Burchinal, 2007-2011). The

NCRECE PDS was a multi-site, randomized trial of two professional development programs designed to enhance the quality of teacher-child interactions in publicly-funded preschool classrooms. The full study included over 400 teachers in over 200 preschool centers in 9 U.S. cities. In the first phase of the study, 427 teachers were randomly assigned to participate in a 14-week course or to a control condition. In the second phase of the study, 401 teachers were randomly assigned to participate in a web-mediated coaching program, or to a control group. Most teachers participated in both phases of the study; therefore, teachers could have participated in the course, the coaching, both interventions, or neither intervention.

For the present study, we used data from the second phase of the study. In addition, our analyses examining the links between teachers' participation in professional development and variation in classroom quality focused only on the coaching. We focused on NCRECE coaching for two reasons. First, as we rely on data from the second phase of the study, this allows us to examine how variation in classroom quality related to teachers' contemporaneous participation in professional development. Second, the use of coaching to build teacher capacity has become widespread in recent years (Egert et al., 2018; Isner et al., 2011). Understanding how coaching affects multiple aspects of classroom quality has implications for ECE policy and practice.

In the coaching, teachers received observation-based analysis and feedback from coaches via web-mediated interactions. This occurred during regular coaching cycles that took place approximately every two weeks throughout the year. In each cycle, teachers videotaped a short lesson or instructional activity focused on language and literacy. Coaches then provided feedback on the videos through the online portal, and highlighted examples of effective teacher-child interactions that met lesson objectives. Teachers and coaches then discussed the feedback and developed a plan to implement future instructional activities. The focus of the coaching

cycles varied across the intervention period. The first three cycles focused on the emotional support domain of the CLASS, the next two cycles focused on classroom organization, and the remaining cycles focused on instructional support (Pianta et al., 2014). Therefore, teachers and coaches tended to focus more on instructional support relative to the other two domains. (The NCRECE coaching condition is explained in detail in Pianta et al. [2014].)

Throughout the school year, teachers in both the coaching and control conditions submitted videos of instruction that were scored using the CLASS. In addition, children's outcomes were assessed at the beginning and end of the school year. Child outcomes included language, literacy, inhibitory control, and self-regulation information collected for four sampled children in each teacher's classroom via direct assessment and teacher-reports. Details of the specific measures used for teacher and child outcomes are described below; details of the timing of the intervention and data collection are presented in Appendix Figure A1.

Analytic Sample. The present study included 278 preschool teachers and 1,214 children from the NCRECE PDS. To be included in the analytic sample, teachers had to have: (i) been randomly assigned to the coaching or control condition in the second phase of the study, (ii) at least two video submissions scored using the CLASS, and (iii) language, literacy, inhibitory control and/or self-regulation outcome information for at least one child in their classroom.

Table 1 describes the analytic sample. The analytic sample included 278 of the 401 teachers (69 percent) who participated in the second phase of the study. Most teachers excluded from the analytic sample had fewer than two CLASS video submissions (n = 120); a small number of teachers had CLASS scores from at least two videos but did not have child outcome information (n = 3). Teachers in the analytic sample were 42.2 years old and had 14.5 years of experience, on average, and most taught in Head Start or public schools. Most teachers taught in

low-income settings; on average, 88 percent of children in their classrooms had income-to-needs ratios below 2.0. Teachers in the analytic sample were comparable to the full sample on most observed characteristics. However, teachers in the analytic sample were more likely to have been assigned to the coaching condition relative to teachers excluded from the analytic sample. Teachers in the analytic sample also had more years of education, on average, were more likely to be White, and were less likely to be Hispanic relative to excluded teachers.

The analytic sample also included 1,214 of the 1,407 children (86 percent) who participated in the second phase of the study. Children in the analytic sample were 4.2 years old, on average. Their mothers had on average obtained 12.7 years of education, and their average income-to-needs ratio was 1.1. Children in the sample were also racially diverse, including 48 percent Black, 34 percent Hispanic, and 11 percent White. Children in the analytic sample were comparable to the full sample on most observable characteristics. However, children in the analytic sample were somewhat older, on average, relative to excluded children.

Measures

Classroom quality. The quality of interactions between teachers and children was measured using the CLASS (Pianta et al., 2008). The CLASS measures 10 dimensions of interactions, each rated on a 7-point scale, which are aggregated to create three domain scores. CLASS dimensions include positive climate, negative climate, teacher sensitivity, and regard for student perspectives dimensions (which form the emotional support domain); behavior management, productivity, and instructional learning formats (which form the classroom organization domain); and concept development, quality of feedback, and language modeling (which form the instructional support domain). In the analytic sample, Cronbach's were .75 (emotional support), .69 (classroom organization), and .87 (instructional support).

CLASS scores were based on 30-minute, teacher-submitted videos of language and literacy activities filmed throughout the school year, after the start of the intervention. Teachers submitted videos approximately every two weeks (for teachers in the coaching condition) or every four weeks (for teachers in the control condition). Each video was divided into two 15-minute video segments, and each segment was scored by two randomly-assigned, trained CLASS observers. Scores were aggregated across all raters to calculate video scores. Prior to coding videos, observers participated in a two-day training session and had to demonstrate acceptable reliability. Specifically, observers scored five video segments using the CLASS and were required to show consistency with master codes (i.e., within one point) for 80 percent of codes. Observers also showed high reliability during ongoing meetings conducted during the intervention period (Pianta et al., 2014). Rater agreement for video segments within the present analytic sample (i.e., the percent of scores within one point) was generally high, between 79 and 99 percent across CLASS dimensions and domains (see Appendix Table A1).

Table 2 presents descriptive statistics on teachers' CLASS scores. On average, teachers submitted approximately eight CLASS videos during the year (mean = 8.2, SD = 4.3, range = [2, 32]). Consistent with the design of the study, which required treatment teachers to submit videos more often than control teachers, teachers in the coaching condition had 10.0 video submissions, on average, relative to 5.7 video submissions for teachers in the control condition. Videos were submitted throughout the school year from August through June (see Appendix Figure A2).

Table 2 also shows that there was more variation in CLASS scores within rather than across teachers. On average, the unadjusted standard deviations of individual teachers' observed CLASS domain scores range from 0.49 (emotional support) to 0.61 (instructional support). Intra-cluster correlations, estimated from unconditional two-level models with CLASS scores nested

within teachers, yield ICCs for the three CLASS domains of 0.18 (instructional support), 0.32 (classroom organization), and 0.33 (emotional support). Consistent with prior research on the CLASS, emotional support and classroom organization scores in our sample were relatively high on average as compared with instructional support scores (5.27 and 5.38 versus 2.33).

Child outcomes. A composite measure of children's early language skills was built from two individual measures. The first measure of children's language was the 168-item Peabody Picture Vocabulary Test (PPVT), a nationally normed measure of children's receptive vocabulary that generally demonstrates acceptable reliability. Second, the 44-item Woodcock-Johnson III Picture Vocabulary Test was used to capture expressive vocabulary by asking children to identify pictured objects, and demonstrates high internal reliability (Woodcock et al., 2001). These two measures were correlated at 0.73 at both the beginning and end of the year. We constructed our language composite by standardizing (i.e., z-scoring) both measures and calculating the average of the two standardized measures. This composite was calculated for children with information for both outcome measures.

A composite measure of children's early literacy skills was also built from two individual measures. The 36-item Test of Preschool Early Literacy (TOPEL) Print Knowledge subtest was used to assess emergent literacy skills. The Print Knowledge subtest measures children's knowledge of the alphabet, written language conventions, and writing form. The 27-item TOPEL Phonological Awareness subtest captured phonological awareness including word awareness and phonemic awareness (Wilson & Longian, 2010). These two measures were correlated at 0.42 and 0.40 at the beginning and end of the year, respectively. As with the language composite, the literacy composite was calculated based on the average of the two standardized literacy measures, and was calculated for children with information on both measures.

Children's inhibitory control was measured using the 16-item Pencil Tap assessment. The Pencil Tap is an adapted version of the peg-tapping task that asks children to tap a pencil once when the assessor taps twice, and vice versa. This assessment has been frequently used and validated in the literature (Blair & Razza, 2007; Diamond & Taylor, 1996). This measure showed high internal consistency at the end of the year (Cronbach's alpha = .92).

Children's self-regulation skills, including children's persistence and engagement in the classroom, were measured using the 24-item Preschool Learning Behaviors Scale (PLBS; McDermott, Leigh, & Perry, 2002). The PLBS is a teacher-reported measure of children's self-regulation skills, including children's competence, motivation, attention/persistence, and attitudes towards learning. The PLBS has shown predictive validity with respect to children's cognitive outcomes (McDermott et al., 2002). Scores were calculated based on the sum across items and showed high internal consistency at the end of the year (Cronbach's alpha = .91).

School, teacher, and child covariates. Covariates were measured at the beginning of the first or second phase of the study. Teacher- and school-level covariates included teacher age, race/ethnicity (i.e., black, white, Hispanic, or other race/ethnicity), years of experience, years of education, whether the teacher taught in a Head Start program, whether the teacher taught in a public school, classroom poverty, treatment status in each phase of the study, whether the teacher was added to the study in the second phase, and site (city) indicators. Child-level covariates included child age, gender, race/ethnicity (i.e., black, white, Hispanic, or other race/ethnicity), mother's years of education, family income-to-needs ratio, and fall pretest score. We selected covariates that may be related to both classroom quality and children's outcomes, and that have been included in prior analyses of the NCRECE PDS.

Analytic Approach

Associations between overall variation in classroom quality and child outcomes. We first examined the associations between overall variation in teachers' classroom quality and child outcomes. To do so, we calculated the standard deviation of each teacher's observed scores in each CLASS domain over the year. To account for the fact that teachers submitted different number of videos, we applied a shrinkage adjustment to the estimated standard deviations of CLASS scores. This adjustment accounts for differences in the number of video submissions across teachers by shrinking the estimates of teacher-specific CLASS score variation for teachers with fewer observations towards the mean across teachers (see Appendix B for details of this procedure). We then used a simple regression-based approach to examine the association between teacher-specific overall variation in quality and child outcomes. Specifically, we estimated a two-level model of the following form for child *i* in a classroom taught by teacher *j*:

Model 1:

(Level 1) $$Y_{ij} = \beta_{0j} + \beta_1 \overrightarrow{X_{ij}} + \epsilon_{ij}$$

(Level 2) $$\beta_{0j} = \pi_0 + \pi_1 AdjSDCLASS_j + \pi_2 AvgCLASS_j + \pi_3 SqrtNumCLASS_j$$

$$+ \pi_4 AvgCLASS_j * SqrtNumCLASS_j + \pi_5 \overrightarrow{T_j} + \pi_6 Treat_j + u_{0j}$$

where Y_{ij} is the outcome for child *i* in a classroom with teacher *j*, $AdjSDCLASS_j$ is the adjusted measure of overall variation in CLASS scores for teacher *j*, and $Treat_j$ is the treatment assignment for teacher *j* in the second phase of the study. We controlled for the average of teachers' observed CLASS scores ($AvgCLASS_j$) and the square root of the number of teachers' video submissions ($SqrtNumCLASS_j$). We also controlled for their interaction to account for the possibility that the association between average CLASS scores and child outcomes might be larger when teachers have more video submissions and, consequently, more precisely estimated CLASS scores. We also included the vector of child-level covariates ($\overrightarrow{X_{ij}}$) and the vector of

teacher- and school-level covariates ($\overrightarrow{T_j}$), described above. The parameter π_1 is the coefficient of interest that captures the association between teacher-specific variability in CLASS scores and child outcomes above and beyond average quality over the year and other measured teacher characteristics. Separate models were estimated for each CLASS domain and child outcome.

To examine whether the association between overall variation in quality and child outcomes varied based on average quality, we also estimated models that included an interaction between teachers' measure of overall variation in quality and average CLASS scores.

Associations between teacher-specific trends in classroom quality and child outcomes. We then examined the associations between teacher-specific trends in quality and child outcomes. To address this aim, we first estimated a series of random effects linear spline models that examined systematic changes in quality over the year. Estimating these models allowed us to decompose overall variation in quality over the year into variation due to teacher specific-trends and instability around these teacher-specific trends. We then extracted estimates of teacher-specific trends in quality, which comprised our key predictors of interest.

Formally, we first estimated models of the following form for a CLASS domain score from a video submitted at time *t* by teacher *j*, where trends in quality are allowed to vary across the fall (August to December), winter *(*January to March), and spring (April to June)*:*

Model 2:

(Level 1) $$CLASS_{tj} = \beta_{0j} + \beta_{1j}Time_t + \beta_{2j}TimeWinter_t + \beta_{3j}TimeSpring_t + \epsilon_{ij}$$

(Level 2) $$\beta_{0j} = \gamma_{00} + \gamma_{01}Treat_j + \gamma_{02}\overrightarrow{T_j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Treat_j + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}Treat_j$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31}Treat_j$$

where $CLASS_{tj}$ is the emotional support, classroom organizaiton, or instructional support score for a video submitted at time *t* by teacher *j*. $Time_t$ is defined as the number of months relative to January 1. $TimeWinter_t$ takes on the value of 0 if the video was submitted in the fall, the number of months since January 1 if the video was submitted in the winter, and the total number of months between January 1 and April 1 if the video was submitted in the spring. $TimeSpring_t$ takes on the value of 0 if the video was submitted in the fall or winter, and the number of months since April 1 if the video was submitted in the spring. $Treat_j$ is an indicator for whether the teacher was assigned to the coaching condition, and we allowed overall trends in the fall, winter, and spring months to vary by treatment status. As above, we controlled for a series of teacher- and school-level covariates ($\overrightarrow{T_j}$). We also include a random teacher intercept (u_{0j}) and a single random time effect (u_{1j}). The inclusion of these random effects allows the changes in CLASS scores over time to vary randomly across teachers and allows us to estimate a fitted trajectory of quality over the year for each teacher. The specific functional form of this model was based on separate work examining how teachers' classroom quality evolves over time.

The Empirical Bayes estimate of the slope of each teacher's fitted growth curve, $RETime_j$, comprised our key predictor of interest. This represents the teacher-specific change in CLASS scores per month, beyond the average change of teachers in the same treatment condition. To clarify interpretation, this estimate was multiplied by 10 such that it could be interpreted as the teacher-specific change in CLASS scores over (roughly) the full school year. We then used a regression-based approach to examine the association between teacher-specific trends in quality and child outcomes. Specifically, we estimated the same two-level model presented above in Model 1, replacing the adjusted measure of overall variation in CLASS scores for teacher *j*, $AdjSDCLASS_j$, with the Empirical Bayes estimate of the random time effect

for teacher j (multiplied by 10), $10 * RETime_j$. Therefore, this model estimates the association between teacher-specific trends in quality and child outcomes, beyond average quality over the year and other measures covariates. The full model is presented in Appendix A. As above, separate models were estimated for each of the CLASS domains and child outcomes. We also estimated exploratory models including an interaction between the measure of teacher-specific trends in quality and average quality.

Associations between instability in classroom quality and child outcomes. We then examined the associations between instability in quality and child outcomes. To address this aim, we first estimated the same random effects linear spline model described above in Model 2. We then obtained an estimate of each teacher's fitted growth curve, based on estimated fixed effects and teacher-specific random effects. This fitted growth captures the extent to which her classroom quality rose and fell over the year. We then calculated residuals from this model as the difference between teachers' observed CLASS scores and the predicted CLASS scores based on their fitted growth curves. Variation in these residuals represents the residual variability in each teacher's CLASS scores around their fitted growth curve, i.e., variation in classroom quality that is not accounted for by systematic increases or decreases in quality over the year. Therefore, we took the standard deviation of these residuals as our measure of teachers' instability in quality. As in our analyses examining the associations between overall variation in quality and child outcomes, we applied an adjustment that shrinks estimates of teacher-specific instability in quality for teachers with fewer observations towards the mean.

We then re-estimated the two-level model presented above in Model 1, replacing the adjusted measure of overall variation in CLASS scores for teacher j, $AdjSDCLASS_j$, with the adjusted measure of instability in CLASS scores for teacher j, $SDResid_j$. Therefore, this model

estimates the association between instability in quality and child outcomes, beyond average quality over the year and other measures covariates. The full model is presented in Appendix A. Separate models were again estimated for each of the CLASS domains and child outcomes. As above, we also estimated exploratory models including an interaction between the measure of instability in quality and average quality.

Associations between participation in coaching and variation in classroom quality. Finally, we examined whether various aspects of quality variation, including overall variation, trends in quality, and instability around teacher-specific trends, differed for teachers assigned to the coaching intervention relative to teachers assigned to the control condition. To do so, we first returned to the random effects linear spline model presented in Model 2. We examined whether average trends in quality in the fall, winter, or spring months differed by teachers' treatment status by examining estimates of γ_{11}, γ_{21}, and γ_{31}. These parameters capture whether the change in CLASS scores per month differed between treatment and control teachers during the fall, winter, and spring, respectively. Next, to examine whether participation in the coaching was associated with overall variation in quality, we used teachers' treatment status to predict the adjusted overall standard deviation of CLASS domain scores, controlling for teacher and school covariates. We estimated similar models examining the associations between coaching and instability in quality. Details of these models are presented in Appendix A.

As noted above, the present analysis includes only a subsample of the teachers who were randomly assigned to the coaching or control conditions at the start of the second phase of the NCRECE study. Moreover, our analytic sample includes a disproportionate number of treatment teachers. We therefore interpret our findings regarding the links between teachers' participation in the coaching and quality variation as correlational rather than causal.

Missing data. In the analytic sample, missingness rates for covariates were generally low (from 0 to 24 percent). Missing covariate information was handled using simple mean imputation with missing variables: missing covariates were set to the sample mean and a dummy variable indicating missing covariate information was included in analyses. We used this approach because our analysis includes extracting Empirical Bayes estimates of teacher-specific trends in quality from our random effects linear spline models. Extracting these estimates is not currently supported when using more complex methods to handle missing data (e.g., multiple imputation).

Results

Associations Between Overall Variation in Classroom Quality and Child Outcomes

Results presented in Table 3 show the associations between overall variation in teachers' CLASS scores and child outcomes. Overall, results provide some evidence that classrooms characterized by higher overall variation in classroom quality were associated with less positive gains in children's early development in terms of inhibitory control, literacy, and self-regulation. Children in classrooms with more overall variation in emotional support had lower inhibitory control. Specifically, a one-point difference in the teacher-specific standard deviation of emotional support scores was associated with a 1.18 SD decline (SE = 0.419, $p < .01$) on inhibitory control. The cross-teacher standard deviation of overall variation in emotional support was 0.08; this suggests children's inhibitory control in a classroom one standard deviation above the mean was 0.09 SD lower relative to children in a typical classroom. Similarly, children in classrooms with more overall variation in classroom organization had lower self-regulation and literacy skills. A one-point difference in teacher-specific standard deviation of classroom organization scores was associated with a 0.79 SD decline (SE = 0.38, $p < .05$) in self-regulation and a 0.58 decline (SE = 0.29, $p < .05$) in literacy. As the cross-teacher standard deviation of

overall variation in classroom organization was 0.11, this indicates that children in a classroom one standard deviation above the mean showed declines of 0.09 SD in self-regulation and 0.06 SD in literacy, relative to children in a typical classroom.

However, associations between variation in other CLASS domains and children's inhibitory control, self-regulation, and literacy outcomes were not statistically significant. Additionally, none of the associations between overall variation in classroom quality and children's language outcomes were statistically significant (Associations between overall variation in classroom quality and individual components of the language and literacy composites are consistent with these findings; see Appendix Table A2). Moreover, associations did not differ based on average quality (see Appendix Table A3).

Across all estimates, associations between overall variation in classroom quality were generally negative in sign but not always precisely estimated enough as to be distinguished from zero. As we have examined 12 primary relationships (four child outcomes for each of the three CLASS domains) we did a further analysis to guard against multiple testing concerns. In particular, we conducted a series of permutation tests to further examine whether the above results taken as a whole indicate a general, overall relationship between variation in classroom quality and child outcomes. Specifically, we permuted the teacher-level triples of variation in classroom quality across teachers such that each teacher was randomly assigned the values of variation in quality of another teacher. The joint permutation of overall variation in emotional support, classroom organization, and instructional support accounts for the correlation structure across the three CLASS domains. For each permutation, we re-estimated our primary models examining the associations between variation in quality and all child outcomes, using the permuted values of the classroom quality measures and the observed values of child outcomes

and all covariates. Our final test statistics were the average association across child outcomes for a specific quality measure. We also used an overall average of averages to give an overall test of the relationship. For each statistic, we compared the observed average association to the distribution of averages from our simulated data. Results of this exercise, presented in more detail in Appendix C, suggest that the observed associations between variation in quality and child outcomes were more negative than we would expect by chance overall ($p \approx 0.020$). The overall association appeared to be primarily driven by emotional support and classroom organization. These tests verify that there was *some* relationship between variation in quality and child outcomes. The subsequently presented investigations examine which aspects of variation – teacher-specific trends in quality or instability around those trends – are driving this relationship.

Associations Between Teacher-Specific Trends in Classroom Quality and Child Outcomes.

Results of the random effects linear spline model indicate that there was variation in teacher-specific trends in classroom quality, with some teachers demonstrating more positive growth over the school year relative to other teachers (Table 4). As shown in Figure 3, some teachers experienced more positive growth in emotional support per month relative to the average in the same treatment condition (e.g., as much as an additional 0.5 points over the year on the 1-7 CLASS scale); other teachers experienced less growth relative to the typical teacher (e.g., 0.5 points less over the year). We observed a similar amount of variation in teacher-specific trends in classroom organization and instructional support (e.g., teachers' changes in quality over the year ranged from an increase of 0.5 points to a decrease of 0.5 points, relative to the typical teacher in the same treatment condition). However, as shown in the top panel of Table 5, there was little association between these teacher-specific trends in quality and child outcomes. Across all CLASS domains, children in classrooms where teachers demonstrated more positive growth

over the year relative to other teachers in the same treatment condition, with the same average CLASS score, did not show greater gains in language, literacy, inhibitory control, or self-regulation (after accounting for average quality).

Associations Between Instability in Classroom Quality and Child Outcomes.

As shown in the lower panel of Table 5, we observed some evidence that instability in quality was negatively associated with children's inhibitory control development, above and beyond growth trends. Specifically, an increase in residual instability of one point in emotional support was associated with a decrease in inhibitory control of 1.01 SD (SE = 0.40, $p < .05$). As the cross-teacher standard deviation in instability in emotional support is 0.08, this indicates that children's inhibitory control was 0.08 SD lower for children in classrooms one standard above the mean, relative to children in a typical classroom. We also found some evidence that residual instability in classroom organization was negatively associated with children's self-regulation and literacy; however, these associations were only marginally significant. Associations between residual instability in classroom quality and child outcomes were smaller in magnitude and not statistically significant for other CLASS domains and child outcomes.

Overall, these results coupled with the null results regarding teacher-specific trends provide some evidence that the negative associations between overall variation in quality and child outcomes were primarily driven by residual instability in quality rather than teacher-specific trends in growth or decline over the year. Associations did not differ based on average levels of quality (see Appendix Table A4).

We conducted a similar permutation test to examine whether the generally negative associations between residual instability in classroom quality and children's outcomes may be driven by a combination of spurious associations and correlated outcomes. For the emotional

support domain, the observed associations were more negative than we would expect to see by chance (see Appendix C for more details). In contrast, for the classroom organization and instructional support domains, results of the simulation exercise suggest that we might observe associations of the magnitudes found in Table 5 by chance.

Associations Between Participation in Coaching and Variation in Classroom Quality.

The results of estimating the random effects linear spline model, described above and shown in Table 4, demonstrate the possible impact of the coaching on trends in classroom quality. Overall, trends in quality for either emotional support or classroom organization did not differ for treatment and control teachers. In contrast, teachers assigned to the coaching condition showed more positive growth in instructional support. This was driven by more positive growth among treatment group teachers relative to teachers in the control group during the winter months of the school year. Results of a likelihood ratio test of the three treatment coefficients indicates that allowing trends in instructional support to vary between treatment and control groups led to a statistically significant improvement in model fit ($p < 0.001$). Although we note that our ability to make causal inferences about this association is limited due to our focus on a subset of teachers included in the larger experimental study, this suggests that the coaching may have impacted trends in quality over the year.

Our findings also suggest that the coaching may have affected overall variation and instability in quality. As shown in Table 6, we did not observe significant differences between treatment and control teachers in the amount of overall variation or instability in quality for either emotional support or classroom organization, although all estimated impacts are positive in sign. However, results indicate that the coaching may have increased variation in instructional support. We observed a difference in overall variation in instructional support between treatment

and control teachers of 0.05 *(SE* = 0.01, $p < .001$). We also observed a difference in residual instability between treatment and control teachers of 0.03 (*SE* = 0.01, $p < .001$).

Discussion

In recent years, researchers have made strides in defining and measuring the features of classroom environments that matter for children, particularly in terms of the quality of interactions between teachers and children (Mashburn et al., 2008; Pianta et al., 2008). This has led to growth of observation-based ECE accountability systems focused on teacher-child interactions (Mashburn, 2017), and to professional development (e.g., coaching) aligned with these measures (Pianta et al., 2008). Yet, most research on and training to improve teacher-child interactions has focused on children's average experiences in the classroom. In research and policy, quality ratings rely largely on point-in-time observations or measures of quality averaged across multiple observations (Hamre et al., 2012; Mashburn et al., 2008; Perlman et al., 2016).

Emerging literature suggests averaging classroom quality or using one observation point may miss variation in ways that have implications for children's learning (Brock & Curby, 2014; Curby et al., 2013). Our study directly tests this possibility. We examined whether three types of quality variation over the preschool year – overall variation in quality, teacher-specific trends in quality, and instability in quality around these trends – related to children's inhibitory control, self-regulation, language, and literacy outcomes. Our results indicate the amount of overall variation in classroom quality – particularly in emotional support and classroom organization – was negatively associated with children's inhibitory control, self-regulation, and literacy. We observed some evidence that these results were driven by associations between instability in quality and child outcomes, rather than trends. Moreover, our results suggest that teachers' participation in coaching increased instability in instructional support. These results shed light on

how classroom environments support children's learning, and highlight considerations for research and policy efforts to improve the quality of ECE classroom environments.

Variation in classroom quality and children's early learning outcomes

Research shows that unstable home environments can inhibit children's early development (Evans, 2006; Martin et al., 2011), and that children thrive when they have consistent, supportive relationships with parents and caregivers (Kohen et al., 2000; Maccoby, 2000). Yet, there is little understanding of the role of stability in early education environments, and in children's relationships with teachers. Our findings suggest that *stability* in the quality of children's early learning environments may be salient for children's development. These findings extend prior studies that found positive associations between within-day stability in emotional support and children's regulatory outcomes (Brock & Curby, 2014; Curby et al., 2014). First, we show that stability in ECE classroom quality *across the year* similarly relates to children's regulatory outcomes. Second, we show that stability in quality in ECE settings may also support children's development of cognitive skills. This is consistent with prior research showing that children's regulatory and cognitive skills are highly correlated, and that effective early interventions can promote both aspects of children's development (Blair & Razza, 2007).

Our findings are somewhat consistent with hypotheses about the between- and within-domain links between quality, as operationalized by the CLASS, and child outcomes (Downer, Sabol, and Hamre, 2010). We observed associations between instability in classroom organization and children's self-regulation skills; prior evidence also showed classroom organization is associated with children's self-regulation (e.g., Rimm-Kaufman et al., 2009). Consistent with prior evidence of cross-domain associations between emotional support and self-regulation (e.g., Pianta et al., 2002) and between classroom organization and academic skills

(e.g., Freiberg, Connell & Lorentz, 2001), we also observed that variation in emotional support was negatively linked with children's inhibitory control and that variation in classroom organization was negatively linked with children's literacy skills.

However, in contrast to research that suggests instructional support is most strongly linked to children's early academic and cognitive outcomes (e.g., Howes et al., 2008; Mashburn et al., 2008), variation in instructional support was not associated with children's language or literacy outcomes in our sample. The fact that instructional support scores were generally low in our sample may explain this null result. Evidence suggests associations between instructional support and child outcomes are stronger at higher ranges of quality (Burchinal et al., 2010). Alternatively, variation in instructional support, as compared with emotional support and classroom organization, may be less salient for children's development. Instability in the amount of warmth and closeness in children's interactions with their teachers, or unpredictability in children's classroom routines or procedures, may disrupt children's ability to build supportive relationships with teachers. Overall, more work is needed to understand the multi-faceted ways in which early learning environments support children's development.

The data used in this study do not contain detailed information on other time-varying aspects of classroom environments or classroom behaviors that might help us better disentangle the mechanisms underlying fluctuations in classroom quality. For example, time-varying aspects of teachers and classrooms may contribute to fluctuations in classroom quality over time, such as changes in teachers' knowledge, beliefs, or stress. Moreover, teachers may be able to provide a more stable classroom environment in settings where children's behavioral or self-regulation skills develop over time. Future research should continue to unpack the factors that contribute to stability, or instability, in classroom quality over time.

Implications for the use of observational measures of quality in ECE accountability policy

Our findings also suggest that researchers and policymakers who seek to identify high-quality ECE classrooms should look beyond average levels of quality. Classrooms with lower average quality, but where quality is more stable, may be more supportive of children's early development as compared with classrooms characterized by higher average quality that is more unstable. Due to the correlational nature of the analyses in our study, we cannot directly examine potential tradeoffs between increasing average quality and increasing instability quality – for example, whether children benefit from being in classrooms with low but stable quality as compared to classrooms with high but variable quality. Yet, we found negative associations between variation and stability in quality and child outcomes at both higher and lower ranges of average classroom quality. Taken together, our findings suggest that such tradeoffs may exist.

Our results point to opportunities to better harness the use of observational measures such as the CLASS in ECE accountability systems, such as in the Head Start DRS and state QRIS. Current state and federal accountability efforts typically rely on a small number of program observations conducted within a relatively short time period within the year (e.g., typically within the same week) given the cost of conducting multiple observations (Office of Head Start, 2016). Our results suggest that these efforts may yield an incomplete picture of children's experiences by not capturing *variation* in quality over the school year. In practice, multiple observations across multiple days, weeks, or months could help identify those settings characterized by both high and stable quality. Investing additional resources in classroom observations in order to be able to take average quality and variation in quality into account, as opposed to having only a snapshot of quality, may help researchers and policymakers better identify those settings that can support children's early development. Future research that takes a

more comprehensive approach to measuring and conceptualizing quality – specifically, research that considers how classroom quality evolves over time– can help shed light on this question and inform the field's efforts to measure and improve quality at scale.

Teachers participation in coaching and variation in classroom quality

Our findings suggest that teachers' participation in coaching not only increased average quality, but may also have led to increased variation and instability in some aspects of quality. Specifically, teachers assigned to participate in the coaching had more overall variation and residual instability in instructional support, relative to control teachers. An explanation for this result is that the focus of the coaching, and its impacts on classroom quality, may have varied over the year. Some research suggests that suggests individual coaching cycles generated short-term quality improvements in classroom quality that did not persist over time (Hanno, 2020). Moreover, the coaching cycles focused on different domains of quality; individual coaching cycles may have generated improvements in quality only within targeted domains. Indeed, Pianta et al. (2014) found that watching videos focused on one domain was negatively associated with growth in quality in other domains. Therefore, changes in coaching cycle focus over the year may also have generated fluctuations in quality as teachers shifted their instructional focus from non-targeted to targeted domains of quality.

These findings also have implications for widespread use of coaching as a professional learning tool. Our results suggest that coaching and other professional development efforts should, at a minimum, focus on raising without destabilizing quality. In practice, professional development efforts could also take a more comprehensive view of teachers' classroom quality by taking both average quality and instability (e.g., the extent to which teachers' observed quality has fluctuated between recent observations) into account when setting improvement

goals. Coaches could provide guidance on how to implement new practices in challenging or disruptive contexts, in order to reduce large swings in quality.

Limitations

We recognize several limitations in our study. First, our analyses are correlational rather than causal. We hypothesize that variation and instability in quality may lead to lower child outcomes. However, it is possible that the directionality of this relation may be reversed. For example, low levels of child inhibitory control and self-regulation may prevent teachers from providing a stable classroom environment. All analyses controlled for a robust set of child and classroom characteristics, including baseline measures of child outcomes. Nevertheless, we cannot rule out the possibility that unobserved classroom characteristics correlated with observed variation in quality may drive our results.

Our measures of variation in classroom quality may also reflect rater effects. Observers were randomly assigned to score videos. Our measures of variation in quality could reflect, in part, teacher videos being assigned to raters who assigned higher or lower scores. However, rater effects are unlikely to explain our findings. First, as raters were randomly assigned to videos, rater effects would be independent of children's skills. Second, CLASS scores from a single video observation generally represented the aggregate of scores assigned by four raters (i.e., scores from two double-coded video segments). Third, results of a variance decomposition analysis that partitioned variation in CLASS scores assigned by individual raters into variation explained by teachers and raters indicate that raters explain a relatively small proportion of variation in CLASS scores (see Appendix Table A5). We note that raters explain more variation in instructional support as compared to the other two CLASS domains. This may be due in part to the truncated range of instructional support scores observed in our data, as compared to the

other two domains, although we note that we expect truncated scores to generally limit variability. Even if rater effects were similar, they may explain a larger fraction of the more limited amount of overall variation in instructional support scores. Overall, rater effects would lead to overestimation and noise in our estimates of variation in classroom quality. As such, our results may underestimate the associations between quality variation and child outcomes.

Additionally, teachers self-selected the videos of instruction that were submitted and scored with the CLASS. Therefore, the quality of instruction in these videos may not have represented teachers' typical practice. Moreover, teachers in the coaching condition may have selected videos that highlighted aspects of their instruction where they felt they could benefit from additional feedback during the coaching cycles. Variation in observed CLASS scores may, in part, reflect teachers' decisions to submit videos of higher- and lower-quality instruction.

The number of CLASS videos also varied across teachers. We took several steps to account for these differences. As described above, we adjusted our estimates of overall variation and instability in quality to account for differences in the number of teachers' video submissions and controlled for the number of video submissions in analyses linking variation in quality to child outcomes. Additionally, although our analytic sample includes classrooms where teachers had CLASS scores from at least two videos, we confirmed results were similar after restricting the sample to classrooms led by teachers with more videos (see Appendix Figures A3 to A6). We also confirmed similar results after adding a series of indicators for the number of teachers' video submissions to these models, which accounts for fixed observed and unobserved differences between teachers with different numbers of video submissions (see Appendix Tables A6 and A7). Nevertheless, having a consistent number of videos per teacher would allow us to better examine the links between variation in quality and children's outcomes.

Finally, we recognize that the generalizability of our findings may be limited. The NCRECE PDS included center based ECE programs serving low income children. Although centers were spread across multiple cities and states, the sample of classrooms was almost entirely comprised of Head Start programs and centers located in public schools. Therefore, findings from our study are likely to generalize best to those settings and may not provide insight into how variation in classroom quality relates to child outcomes in other ECE contexts.

Conclusions

Decades of research have highlighted the critical role teachers play in children's development. The present study demonstrates that variation in classroom quality across the school year, above and beyond average quality, is associated with children's inhibitory control and self-regulation. Our findings suggest that stable classroom environments – particularly those that provide consistent levels of emotional support and classroom organization – may be important for children's learning. Based on our findings, current approaches to identifying high quality classrooms in both research and policy, which often rely on snapshots of classroom quality based on observations conducted over a short period of time, can be improved. This study also indicates that efforts to improve classroom quality, such as through coaching interventions, should not only aim to improve overall levels of classroom quality, but also support teachers to develop classroom environments that are stable over time. Understanding the full picture of classroom quality – including the ups and downs of quality over the year – matters for ECE programs, policy, and practice.

References

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child development*, *78*(2), 647-663. doi: 10.1111/j.1467-8624.2007.01019.x.

Bronfenbrenner, U. (1979). *The ecology of human development*. Harvard University Press.

Bronfenbrenner, U., & Morris, P. A. (1998). The ecology of developmental processes.

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. doi: 10.1080/10409289.2014.866020

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. doi: 10.1080/03004430.2016.1178245

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

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. doi: 10.1016/j.ecresq.2009.10.004

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

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C.

(2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757-783. doi: 10.1177/0013164413486987

Cassidy, D. J., Hestenes, L. L., Hansen, J. K., Hegde, A., Shim, J., & Hestenes, S. (2005). Revisiting the two faces of child care quality: Structure and process. *Early Education and Development, 16*(4), 505-520. doi: 10.1207/s15566935eed1604_10

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. doi: 10.1080/10409289.2012.665760

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). Effective teacher professional development.

Derrick-Mills, T., Burchinal, M., Peters, H. E., De Marco, A., Forestieri, N., Fyffe, S., ... & Triplett, T. (2016). Early Implementation of the Head Start Designation Renewal System. OPRE Report: 2016-75a. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services.

Downer, J., 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 and Development, 21*(5), 699-723.

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. doi: 10.1016/j.ecresq.2016.08.005

Egert, F., Fukkink, R. G., & Eckhardt, A. G. (2018). Impact of in-service professional development programs for early childhood teachers on quality ratings and child outcomes: A meta-analysis. *Review of Educational Research*, *88*(3), 401-433. doi: 10.3102/0034654317751918

Evans, G. W. 2006. Child development and the physical environment. *Annual Review of Psychology*, 57: 423–451.

Freiberg, H. J., Connell, M. L., & Lorentz, J. (2001). Effects of Consistency Management® on student mathematics achievement in seven chapter I elementary schools. *Journal of Education for Students Placed at Risk*, *6*(3), 249-270. doi: 10.1207/S15327671ESPR0603_6

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher–child interactions: Associations with preschool children's development. *Child Development*, *85*(3), 1257-1274. doi: 10.1111/cdev.12184

Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., LoCasale-Crouch, J., Downer, J. T., ... & Scott-Little, C. (2012). A course on effective teacher-child interactions: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal*, *49*(1), 88-123. doi: 10.3102/0002831211434596

Hanno, E.C. (2020). Short-term changes, trade-offs, and fadeout in early educator practices during coaching. In preparation.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale*. Teachers College Press, Columbia University, New York: NY.

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. doi: 10.1016/j.ecresq.2007.05.002

Isner, T., Tout, K., Zaslow, M., Soli, M., Quinn, K., Rothenberg, L., & Burkhauser, M. (2011). Coaching in Early Care and Education Programs and Quality Rating and Improvement Systems (QRIS): Identifying Promising Features. Washington, DC: Child Trends.

Keys, T. D., Farkas, G., Burchinal, M. R., Duncan, G. J., Vandell, D. L., Li, W., … Howes, C. (2013). Preschool Center Quality and School Readiness: Quality Effects and Variation by Demographic and Child Characteristics. *Child Development*, *84*(4), 1171–1190. doi: 10.1111/cdev.12048

Kohen, D. E., Leventhal, T., Dahinten, V. S., & McIntosh, C. N. (2008). Neighborhood disadvantage: Pathways of effects for young children. *Child Development*, *79*(1), 156-169. doi: 10.1111/j.1467-8624.2007.01117.x

Maccoby, E. E. (2000). Parenting and its effects on children: On reading and misreading behavior genetics. *Annual Review of Psychology*, *51*(1), 1-27. doi: 10.1146/annurev.psych.51.1.1

Malmberg, L. E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology*, *102*(4), 916. doi: 10.1037/a0020920

Malmberg, L. E., Hagger, H., & Webster, S. (2014). Teachers' situation-specific mastery experiences: teacher, student group and lesson effects. *European Journal of Psychology of Education*, *29*(3), 429-451. doi: 10.1007/s10212-013-0206-1

Martin, A., Razza, R. A., & Brooks-Gunn, J. (2012). Specifying the links between household chaos and preschool children's development. *Early child development and care*, *182*(10),

1247-1263. doi: 10.1080/03004430.2011.605522

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

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

Meyer, J. P., Cash, A. H., & Mashburn, A. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment, 16*(4), 227-243. doi:10.1080/10627197.2011.638884

Office of Head Start. (2016). *Designation Renewal System: DRS by the Numbers*. Washington, DC: Office of Head Start, Office of the Administration for Children and Families.

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*(12), 1–33. doi: 10.1371/journal.pone.0167660

Pianta, R. C., & Burchinal, M. (2007–2011). *National center for research on early childhood education professional development study*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor], 2016-04-12. doi:10.3886/ICPSR34848.v2

Pianta, R. C., DeCoster, J., Cabell, S., Burchinal, M., Hamre, B. K., Downer, J., ... & Howes, C. (2014). Dose-response relations between preschool teachers' exposure to components of

professional development and increases in quality of their interactions with children. *Early Childhood Research Quarterly, 29*(4), 499-508. doi: 10.1016/j.ecresq.2014.06.001

Pianta, R., Hamre, B., Downer, J., Burchinal, M., Williford, A., Locasale-Crouch, J., ... & Scott-Little, C. (2017). Early childhood professional development: Coaching and coursework effects on indicators of children's school readiness. *Early Education and Development*, *28*(8), 956-975. doi: 10.1080/10409289.2017.1319783

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

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal*, *102*(3), 225-238.

Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *23*(4), 431-451. doi: 10.1016/j.ecresq.2008.02.001

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. doi: 10.1016/j.learninstruc.2013.12.002

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

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., & Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Developmental Psychology, 45*(4), 958.

doi: 10.1037/a0015861

Sabol, T. J., & Pianta, R. C. (2012). Recent trends in research on teacher–child relationships. *Attachment & Human Development*, *14*(3), 213-231. doi: 10.1080/14616734.2012.672262

Sabol, T. J., Ross, E. C., & Frost, A. (2020). Are All Head Start Classrooms Created Equal? Variation in Classroom Quality Within Head Start Centers and Implications for Accountability Systems. *American Educational Research Journal*, *57*(2), 504-534. doi: 10.3102/0002831219858920

Sandilos, L. E., Goble, P., Rimm-Kaufman, S. E., & Pianta, R. C. (2018). Does professional development reduce the influence of teacher stress on teacher–child interactions in pre-kindergarten classrooms?. *Early Childhood Research Quarterly*, *42*, 280-290. doi: 10.1016/j.ecresq.2017.10.009

Vandenbroucke, L., Spilt, J., Verschueren, K., Piccinin, C., & Baeyens, D. (2018). The classroom as a developmental context for cognitive development: A meta-analysis on the importance of teacher–student interactions for children's executive functions. *Review of Educational Research*, *88*(1), 125-164. doi: 10.3102/0034654317743200

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. doi: 10.1016/j.ecresq.2014.05.010

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. doi: 10.1016/j.ecresq.2012.12.002

Table and Figures

Table 1. Teacher and child characteristics in the full and analytic sample

Teacher characteristics	Full sample (N = 401)		Analytic sample (N = 278)		Diff. between analytic sample, excluded sample	p-value of difference
	Mean	SD	Mean	SD		
Teacher age	42.34	10.62	42.21	10.99	-0.43	0.723
Teacher race/ethnicity						
Black	0.46	0.50	0.48	0.5	0.06	0.295
White	0.31	0.46	0.34	0.47	0.10	0.060
Hispanic	0.15	0.36	0.12	0.32	-0.10	0.012
Years of experience	14.16	9.13	14.49	9.5	1.17	0.267
Years of education	15.74	1.65	15.87	1.6	0.45	0.016
Head Start	0.57	0.50	0.57	0.5	-0.06	0.491
Public school	0.37	0.48	0.38	0.49	0.07	0.395
Classroom poverty	0.88	0.21	0.88	0.21	0.02	0.640
Coaching (Phase 2)	0.51	0.50	0.58	0.49	0.22	<0.001
Course (Phase 1)	0.38	0.49	0.39	0.49	0.04	0.460
Added in Phase 2	0.17	0.38	0.19	0.39	0.05	0.233
Child characteristics	Full sample (N = 1,407)		Analytic sample (N = 1,214)		Diff. between analytic sample, excluded sample	p-value of difference
	Mean	SD	Mean	SD		
Child age	4.17	0.47	4.18	0.46	0.11	0.002
Child gender: Male	0.49	0.5	0.49	0.5	0.00	0.959
Black	0.47	0.5	0.48	0.50	0.06	0.155
White	0.11	0.32	0.11	0.32	-0.02	0.443
Hispanic	0.34	0.47	0.34	0.47	0.01	0.708
Mother's years of education	12.73	2.04	12.71	1.98	-0.20	0.215
Income-to-needs ratio	1.09	1.01	1.07	0.99	-0.14	0.089
Fall pretest						
PPVT	86.03	16.85	85.87	17	-1.15	0.412
Woodcock Johnson Picture Vocabulary	96.07	16.8	96.14	16.93	0.51	0.715
TOPEL Print Knowledge	96.02	14.61	96.1	14.7	0.59	0.627
TOPEL Phonological Awareness	89.94	13.98	89.9	14.04	-0.29	0.805
Pencil Tap	7.36	5.23	7.38	5.19	0.20	0.650
PLBS	37.25	8.43	37.35	8.54	1.06	0.254

Note: p-values based on t-test comparing teachers/children included in the analytic sample with teachers/children excluded from the analytic sample.

Table 2. Summary statistics of CLASS scores

	N	Mean	SD	Min, Max
CLASS scores from all video submissions				
Emotional Support	2,278	5.27	0.62	2.13, 6.88
Classroom Organization	2,278	5.38	0.66	2.33, 7.00
Instructional Support	2,278	2.33	0.78	1.00, 6.00
Average CLASS score across teachers' video submissions				
Emotional Support				
All teachers	278	5.20	0.42	3.88, 6.25
Treatment	161	5.27	0.44	3.88, 6.25
Control	117	5.12	0.38	4.02, 5.90
Classroom Organization				
All teachers	278	5.31	0.47	3.00, 6.23
Treatment	161	5.30	0.48	3.00, 6.13
Control	117	5.33	0.44	3.97, 6.23
Instructional Support				
All teachers	278	2.23	0.42	1.29, 3.45
Treatment	161	2.33	0.44	1.29, 3.45
Control	117	2.08	0.35	1.38, 2.89
Unadjusted SD of CLASS scores across teachers' video submissions				
Emotional Support	278	0.49	0.20	0.00, 1.21
Classroom Organization	278	0.51	0.23	0.00, 1.41
Instructional Support	278	0.61	0.25	0.06, 1.24
Adjusted SD of CLASS scores across teachers' video submissions				
Emotional Support	278	0.51	0.08	0.36, 0.88
Classroom Organization	278	0.55	0.11	0.37, 1.05
Instructional Support	278	0.67	0.06	0.57, 0.93
Number of video submissions				
All teachers	278	8.19	4.26	2, 32
Treatment	161	10.02	4.46	2, 32
Control	117	5.68	2.20	2, 9
ICC from unconditional two-level model	ICC			
Emotional Support	0.33			
Classroom Organization	0.32			
Instructional Support	0.18			

Note: SDs of CLASS scores were adjusted to account for differences in the number of CLASS videos submitted by teachers. Details of the adjustment procedures are in Appendix B. ICCs based on the results of estimating unconditional two-level models with CLASS domain scores nested within teachers.

Table 3. Associations between child outcomes and overall variation in classroom quality

Adj. overall SD:	Inhibitory control (N = 1,053)	Persistence/ Engagement (N = 910)	Language (N = 1,047)	Literacy (1,029)	p-value of permutation test
Emotional Support	-1.175**	0.041	-0.110	-0.441	0.072
	(0.419)	(0.531)	(0.312)	(0.365)	
Classroom Organization	0.170	-0.794*	0.067	-0.582*	0.076
	(0.333)	(0.380)	(0.244)	(0.286)	
Instructional Support	-0.351	-0.240	-0.209	-0.421	0.262
	(0.608)	(0.697)	(0.442)	(0.524)	
Average across CLASS domains	--	--	--	--	0.020

Note: Standard errors in parentheses. Each cell represents the result of a separate regression. All outcomes were z-scored. All models include child- and teacher/school-level covariates listed in Table 1 and site fixed effects. Details of the permutation test are described in Appendix C.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4. Results of random effects linear spline model characterizing classroom quality over time

	Emotional Support	Classroom Organization	Instructional Support
Fixed effects			
Time (months)	-0.064**	0.018	-0.045+
	(0.022)	(0.024)	(0.027)
Time: Winter	0.121**	-0.044	0.185***
	(0.044)	(0.047)	(0.053)
Time: Spring	0.063	-0.075	0.031
	(0.063)	(0.068)	(0.076)
Treatment	0.118+	-0.018	-0.005
	(0.066)	(0.072)	(0.074)
Treatment*Time	0.003	0.010	-0.021
	(0.027)	(0.029)	(0.032)
Treatment*			
Time: Winter	0.019	0.012	0.241***
	(0.052)	(0.057)	(0.063)
Treatment*			
Time: Spring	-0.000	0.015	-0.128
	(0.070)	(0.075)	(0.084)
Random effects			
SD(Time)	0.039	0.032	0.042
	(0.006)	(0.009)	(0.008)
SD(Teacher)	0.301	0.333	0.258
	(0.019)	(0.022)	(0.020)
Corr(Time,			
Teacher)	0.674	0.935	0.844
	(0.164)	(0.263)	(0.192)
SD(Residual)	0.499	0.543	0.611
	(0.008)	(0.009)	(0.010)
Observations	2278	2278	2278
Results of likelihood ratio test	0.717	0.605	<0.001

Note: Standard errors in parentheses. All models include teacher/school-level covariates listed in Table 1 and site fixed effects. Likelihood ratio test compared fit of models with and without time trends that varied by teacher treatment status. Models were estimated using restricted maximum likelihood estimation (REML). For the likelihood ratio tests, models were re-estimated using maximum likelihood estimation (MLE). Due to convergence issues, random effects were assumed to be independent for classroom organization models estimated using MLE.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5. Associations between child outcomes and variation in quality due to teacher-specific trends and instability around teacher-specific trends

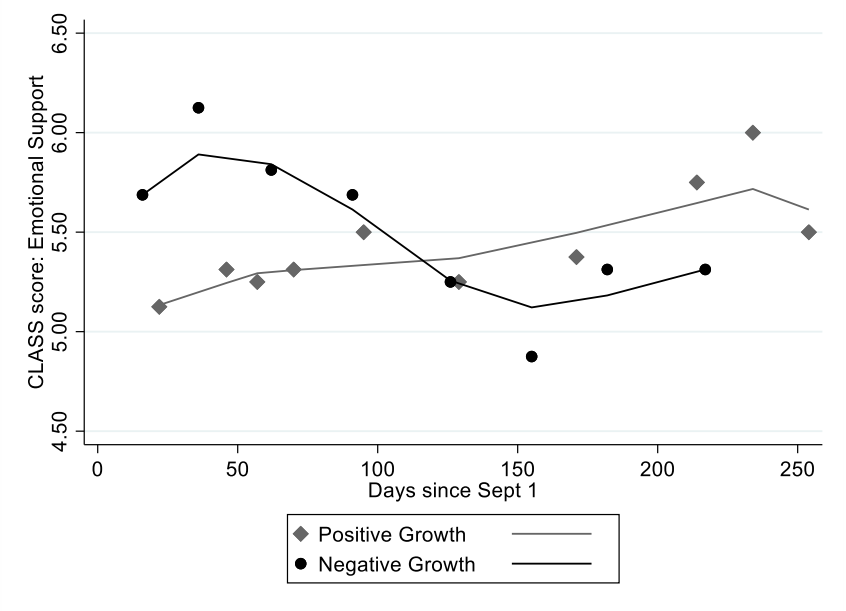
	Inhibitory control (N = 1,053)	Persistence/ Engagement (N = 910)	Language (N = 1,016)	Language (N = 1,016)	p-value of permutation test
Variation:	Teacher-specific trends in quality (growth over the school year)				
Emotional Support	0.157	-0.075	0.076	0.267	--
	(0.263)	(0.313)	(0.193)	(0.226)	
Classroom Organization	-0.228	0.804	-0.223	0.273	--
	(0.661)	(0.804)	(0.482)	(0.575)	
Instructional Support	0.290	0.116	-0.037	0.132	--
	(0.251)	(0.302)	(0.183)	(0.217)	
Variation:	Adj. residual variation around teacher-specific trends in quality				
Emotional Support	-1.005*	-0.082	-0.062	-0.448	0.073
	(0.402)	(0.503)	(0.298)	(0.349)	
Classroom Organization	0.265	-0.707^{+}	0.054	-0.456^{+}	0.179
	(0.318)	(0.363)	(0.234)	(0.274)	
Instructional Support	-0.370	-0.075	0.103	-0.421	0.449
	(0.506)	(0.596)	(0.368)	(0.436)	
Average across CLASS domains	--	--	--	--	0.056

Note: Standard errors in parentheses. Each cell represents the result of a separate regression. All outcomes were z-scored. All models include child- and teacher/school-level covariates listed in Table 1 and site fixed effects. Details of the permutation test are described in Appendix C.
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

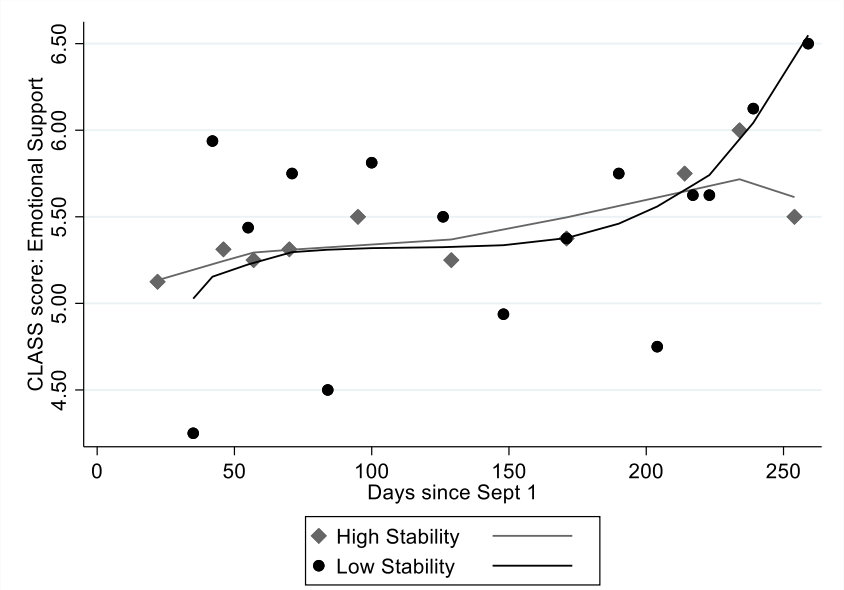
Table 6. Impact of assignment to NCRECE coaching on overall variation in classroom quality and instability in quality around teacher-specific trends

	Emotional Support		Classroom Organization		Instructional Support	
	Adj. overall SD	Adj. residual SD	Adj. overall SD	Adj. residual SD	Adj. overall SD	Adj. residual SD
Coaching	0.004	0.002	0.014	0.015	0.045***	0.031***
	(0.010)	(0.010)	(0.014)	(0.015)	(0.007)	(0.008)
Observations	278	278	278	278	278	278

Note: Standard errors in parentheses. Each cell represents the result of a separate regression. All models include teacher/school-level covariates listed in Table 1 and site fixed effects.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



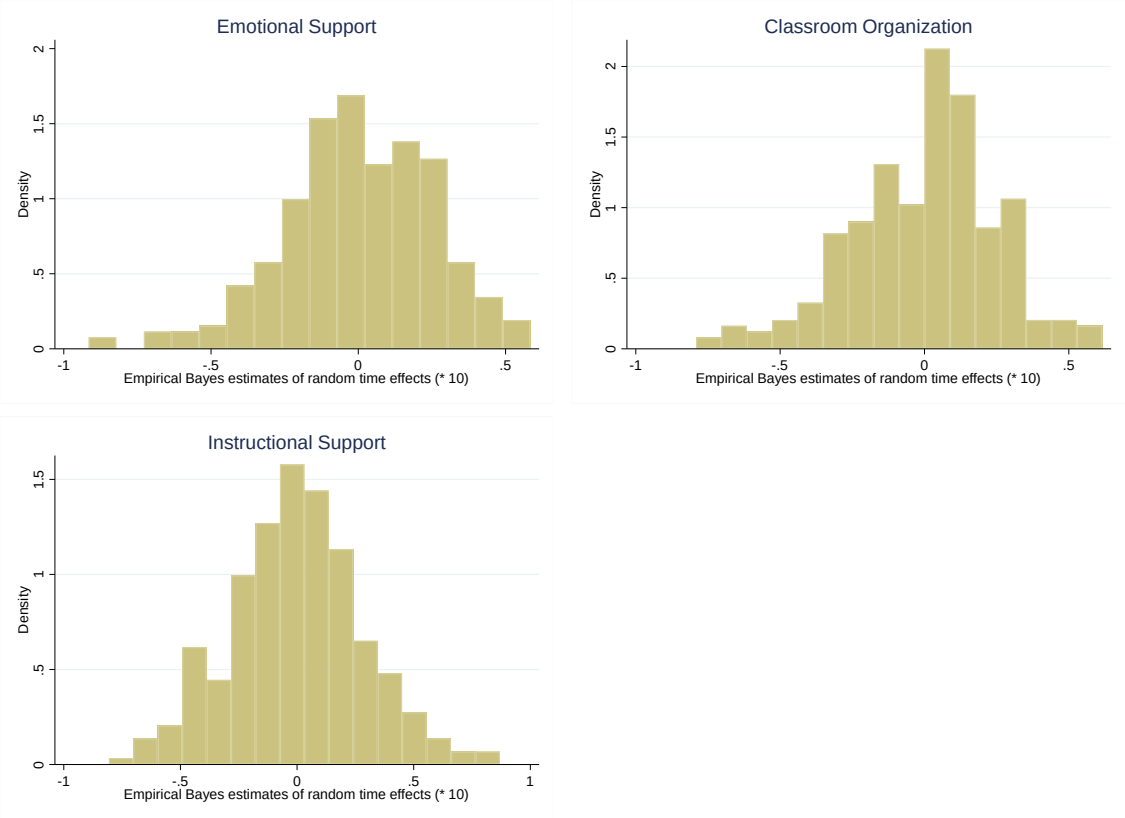
a) Positive and negative trajectories of growth over time



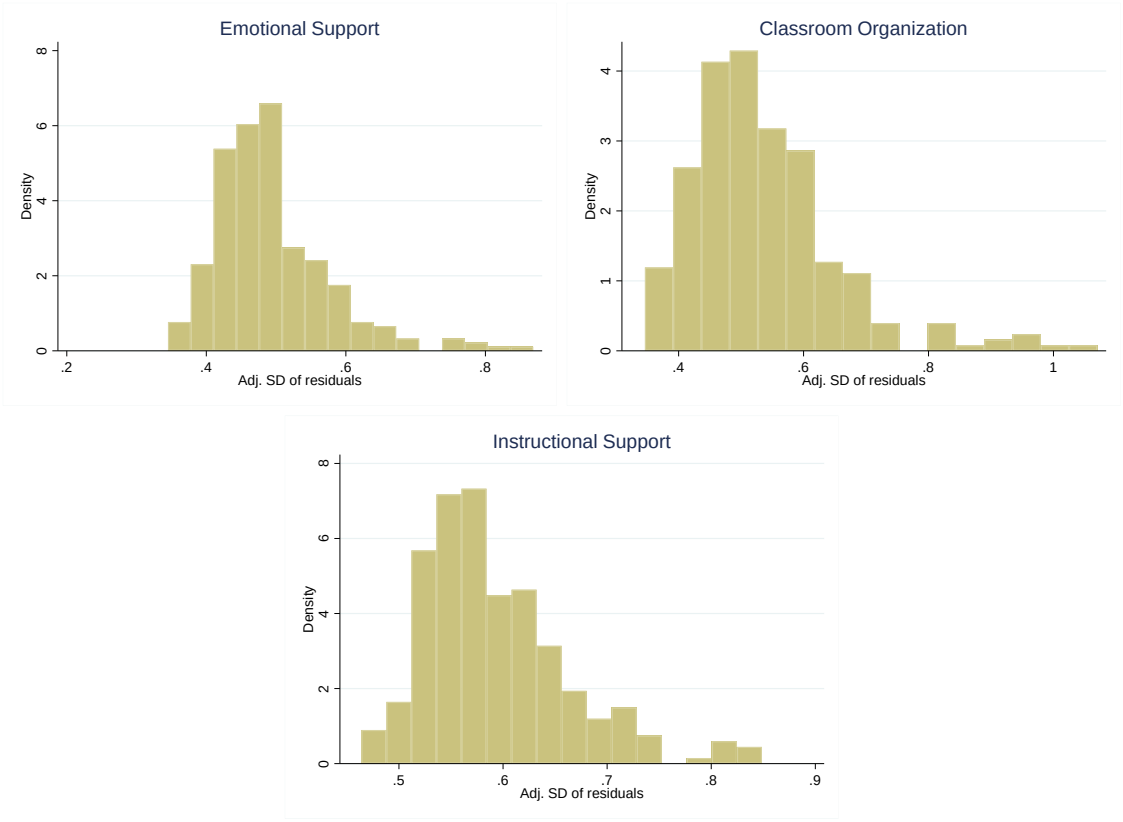
b) High instability and low instability in quality around teacher-specific trends

Figure 1. Teacher trajectories of growth: Positive and negative over time.

Note: Figures present observed emotional support scores and lowess curves of emotional support scores over time. *Figure 1a*: For the "Positive Growth" and "Negative Growth" teachers, average scores were 5.44 and 5.51, respectively. *Figure 1b*: For the "Low Stability" teacher, the average score was 5.44, the teacher-specific slope from an OLS regression of only the teachers' scores on time was $b = 0.002$, and the unadjusted SD of observed scores was 0.26. For the "High Stability" teacher, the average score was 4.46, the teacher-specific slope was $b = 0.003$, and the unadjusted SD of observed scores was 0.61.



a) Empirical Bayes estimates of teacher-specific random time effects, multiplied by 10



b) Adjusted teacher-specific residual standard deviations

Figure 2. Distribution of teacher-specific trends in quality and instability in quality

Note: Teacher-specific random time effects multiplied by 10 to support interpretation as the change in quality over approximately one school year. Teacher-specific residual standard deviations adjusted for teachers' number of video submissions.

Appendix A. Supplemental Tables and Figures

Interrater agreement for CLASS dimension and domain scores

Table A1. Interrater agreement for CLASS dimension and domain scores

	Percent agreement: +/- 1		
	Video segment 1 (*N* = 2,145)	Video segment 2 (*N* = 1,570)	All video segments (*N* = 3,715)
Positive climate	87	87	87
Negative climate	98	99	99
Teacher sensitivity	79	81	80
Regard for student perspectives	79	79	79
Emotional support	*88*	*90*	*89*
Behavior management	88	87	88
Productivity	89	88	88
Instructional learning formats	81	82	81
Classroom organization	*88*	*88*	*88*
Concept development	89	90	89
Quality of feedback	84	87	85
Language modeling	83	86	84
Instructional support	*83*	*84*	*83*

Note: Percent agreement for CLASS domains calculated based on whether domain scores, calculated based on the average score across the relevant dimensions, were within +/- 1 point or in exact agreement. Includes 278 teachers, 2,278 video submissions, and 3,715 video segments.

Table A2. Associations between child outcomes and variation in classroom quality

			Components of Language and Literacy Composite			
Adj. overall SD:	Inhibitory control (N = 1,053)	Persistence/ Engagement (N = 910)	PPVT (N = 1,055)	Woodcock Johnson Picture Vocabulary (N = 1,055)	TOPEL Phonological Awareness (N = 1,032)	TOPEL Print Knowledge (N = 1,055)
Variation: Adjusted overall SD						
Emotional Support	-1.175**	0.041	-0.523	0.101	0.028	-0.715*
	(0.419)	(0.531)	(0.322)	(0.335)	(0.416)	(0.346)
Classroom Organization	0.170	-0.794*	0.126	-0.071	-0.365	-0.725**
	(0.333)	(0.380)	(0.251)	(0.263)	(0.326)	(0.269)
Instructional Support	-0.351	-0.240	-0.259	0.006	-0.388	-0.205
	(0.608)	(0.697)	(0.463)	(0.475)	(0.597)	(0.500)
Variation: Teacher-specific trends in quality (growth over the school year)						
Emotional Support	0.156	-0.074	-0.083	0.262	0.221	0.318
	(0.261)	(0.310)	(0.199)	(0.205)	(0.256)	(0.213)
Classroom Organization	-0.229	0.801	-0.823+	0.419	0.069	0.439
	(0.656)	(0.798)	(0.492)	(0.517)	(0.648)	(0.537)
Instructional Support	0.288	0.114	-0.118	0.129	-0.039	0.332
	(0.248)	(0.299)	(0.189)	(0.194)	(0.244)	(0.203)
Adj. residual variation around teacher-specific trends in quality						
Emotional Support	-1.005*	-0.082	-0.430	0.110	-0.073	-0.631+
	(0.402)	(0.503)	(0.308)	(0.321)	(0.398)	(0.331)
Classroom Organization	0.265	-0.707+	0.155	-0.102	-0.263	-0.596*
	(0.318)	(0.363)	(0.239)	(0.251)	(0.311)	(0.258)
Instructional Support	-0.370	-0.074	0.272	-0.057	-0.446	-0.215
	(0.506)	(0.597)	(0.385)	(0.396)	(0.497)	(0.416)

Note: Standard errors in parentheses. Each cell represents the result of a separate regression. All outcomes were z-scored. All models include child- and teacher/school-level covariates listed in Table 1, and site fixed effects. The first two columns replicate the results in Table 3 and Table 5. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A3. Associations between child outcomes and overall variation in classroom quality, including moderation based on average quality

	Inhibitory control (*N* = 1,053)	Persistence/ Engagement (*N* = 910)	Language (*N* = 1,047)	Literacy (1,029)
Emotional Support:				
Adj. overall SD	-1.347**	-0.023	-0.263	-0.586
	(0.453)	(0.568)	(0.336)	(0.396)
Adj. overall SD*Mean quality	-0.896	-0.406	-0.800	-0.735
	(0.894)	(1.270)	(0.659)	(0.779)
Classroom Organization:				
Adj. overall SD	0.047	-1.006*	0.021	-0.721*
	(0.368)	(0.420)	(0.269)	(0.317)
Adj. overall SD*Mean quality	-0.470	-0.891	-0.175	-0.529
	(0.602)	(0.759)	(0.439)	(0.521)
Instructional Support:				
Adj. overall SD	-0.844	0.033	-0.748	-0.933
	(0.722)	(0.853)	(0.523)	(0.624)
Adj. overall SD*Mean quality	1.383	-0.711	1.510+	1.418
	(1.101)	(1.273)	(0.797)	(0.947)

Note: Standard errors in parentheses. Each cell represents the result of a separate regression. All outcomes were z-scored. All models include child- and teacher/school-level covariates listed in Table 1 and site fixed effects.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A4. Associations between child outcomes and teacher-specific trends in classroom quality, including moderation based on average quality

	Inhibitory control (N = 1,053)	Persistence/ Engagement (N = 910)	Language (N = 1,047)	Literacy (1,029)
Variation:	Teacher-specific trend in quality			
Emotional Support:				
Teacher-specific trend	0.186	-0.042	0.085	0.273
	(0.266)	(0.316)	(0.195)	(0.229)
Teacher-specific trend*Mean quality	0.180	0.226	0.053	0.040
	(0.232)	(0.300)	(0.169)	(0.199)
Classroom Organization:				
Teacher-specific trend	-0.235	0.836	-0.226	0.265
	(0.663)	(0.805)	(0.484)	(0.576)
Teacher-specific trend*Mean quality	-0.071	0.297	-0.031	-0.109
	(0.215)	(0.294)	(0.157)	(0.186)
Instructional Support:				
Teacher-specific trend	0.288	0.116	-0.045	0.126
	(0.251)	(0.303)	(0.183)	(0.217)
Teacher-specific trend*Mean quality	0.070	-0.008	0.222	0.171
	(0.237)	(0.273)	(0.172)	(0.204)
Variation:	Adj. residual variation around teacher-specific trends in quality			
Emotional Support:				
Adj. residual variation	-1.122**	-0.087	-0.214	-0.576
	(0.430)	(0.532)	(0.317)	(0.374)
Adj. residual variation*Mean quality	-0.671	-0.031	-0.876	-0.710
	(0.870)	(1.229)	(0.638)	(0.754)
Classroom Organization:				
Adj. residual variation	0.207	-0.827*	0.026	-0.572+
	(0.353)	(0.405)	(0.258)	(0.304)
Adj. residual variation*Mean quality	-0.216	-0.489	-0.103	-0.434
	(0.571)	(0.733)	(0.416)	(0.495)
Instructional Support:				
Adj. residual variation	-0.768	-0.064	-0.056	-0.773
	(0.578)	(0.691)	(0.422)	(0.498)
Adj. residual variation*Mean quality	1.351	-0.034	0.543	1.193
	(0.962)	(1.099)	(0.702)	(0.828)

Note: Standard errors in parentheses. Each cell represents the result of a separate regression. All outcomes were z-scored. All models include child- and teacher/school-level covariates listed in Table 1 and site fixed effects. + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Variance decomposition of CLASS scores including teacher and rater effects

Results of a variance decomposition indicate that a relatively small amount of variation in teachers' CLASS scores is attributable to rater effects. In our main analysis, CLASS scores for each video submission were based on averaging the scores provided by two (or more) raters. In this variance decomposition exercise, we examine the CLASS scores provided by each rater. Using CLASS scores provided by each rater for each video as the outcome of interest, we estimated a random effects model including crossed random effects for teachers and raters. This analysis includes 7,002 CLASS scores based on 2,278 video submissions.

The percent of the total variation explained by was 7.5 percent (emotional support), 14.2 percent (classroom organization), and 26.5 percent (instructional support). With the exception of instructional support, more variation is explained by teachers. The percent of total variation explained by teachers ranges from 24.1 percent (emotional support), 21.0 percent (classroom organization), and 14.3 percent (instructional support). However, the majority of variation is explained by neither teachers nor raters across the three CLASS domains.

Table A5. Variance decomposition of CLASS scores from individual video submissions.

	Emotional Support	Classroom Organization	Instructional Support
Var(Teachers)	0.141	0.151	0.144
Var(Raters)	0.044	0.102	0.268
Var(Residual)	0.401	0.467	0.598
Total variation	0.586	0.720	1.010
% total variation explained by teachers	24.1%	21.0%	14.3%
% total variation explained by raters	7.5%	14.2%	26.5%

Table A6. Associations between child outcomes and overall variation in classroom quality, including fixed effects for the number of video submissions

Adj. overall SD:	Inhibitory control (N = 1,053)	Persistence/ Engagement (N = 910)	Language (N = 1,047)	Literacy (1,029)
Emotional Support	-1.089*	-0.124	-0.194	-0.498
	(0.455)	(0.572)	(0.333)	(0.389)
Classroom Organization	0.154	-0.975*	-0.118	-0.533+
	(0.344)	(0.392)	(0.251)	(0.292)
Instructional Support	-0.560	-0.135	-0.185	-0.624
	(0.631)	(0.713)	(0.455)	(0.539)

Note: Standard errors in parentheses. Each cell represents the result of a separate regression. All outcomes were z-scored. All models include child- and teacher/school-level covariates listed in Table 1 and site fixed effects. All models also included a series of indicators for the number of teacher video submissions, including 1, 2, 3, … 32 videos.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A7. Associations between child outcomes and variation in quality due to teacher-specific trends and instability around teacher-specific trends, including fixed effects for the number of video submissions

	Inhibitory control (N = 1,053)	Persistence/ Engagement (N = 910)	Language (N = 1,016)	Literacy (1,029)
Variation:	Teacher-specific trends in quality (growth over the school year)			
Emotional Support	0.098	0.029	0.182	0.164
	(0.270)	(0.325)	(0.196)	(0.231)
Classroom Organization	0.401	1.269^{+}	0.560	0.509
	(0.511)	(0.656)	(0.369)	(0.437)
Instructional Support	0.217	0.171	0.006	0.509
	(0.242)	(0.291)	(0.175)	(0.437)
Variation:	Adj. residual variation around teacher-specific trends in quality			
Emotional Support	-0.909*	-0.190	-0.133	-0.527
	(0.434)	(0.538)	(0.317)	(0.369)
Classroom Organization	0.241	-0.883*	-0.130	-0.411
	(0.328)	(0.372)	(0.239)	(0.279)
Instructional Support	-0.553	0.127	0.079	-0.425
	(0.524)	(0.613)	(0.379)	(0.448)

Note: Standard errors in parentheses. Each cell represents the result of a separate regression. All outcomes were z-scored. All models include child- and teacher/school-level covariates listed in Table 1 and site fixed effects. All models also included a series of indicators for the number of teacher video submissions, including 1, 2, 3, … 32 videos.
$+ p < 0.10$, $^{*} p < 0.05$, $^{**} p < 0.01$, $^{***} p < 0.001$.

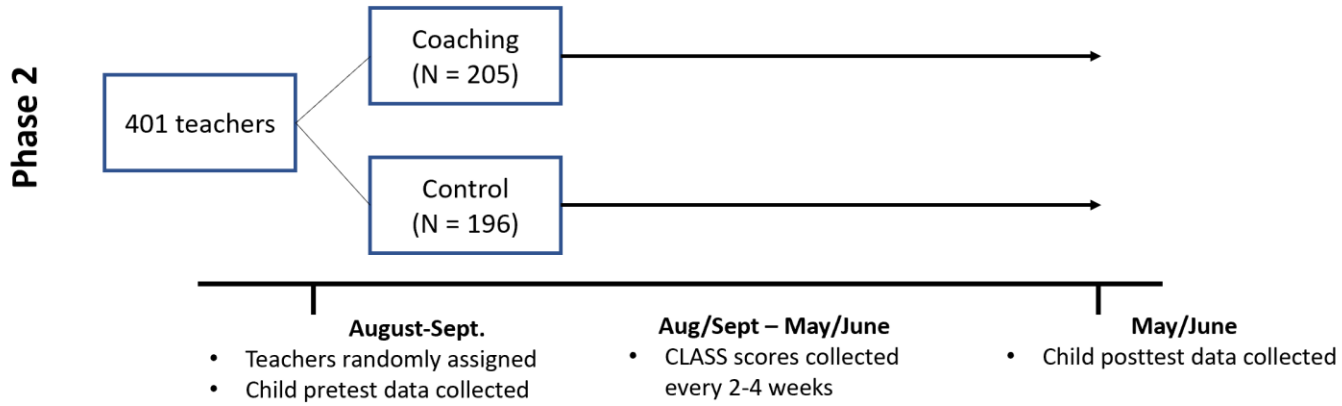


Figure A1. Timing of the NCRECE coaching intervention in the second phase of the NCRECE Professional Development Study

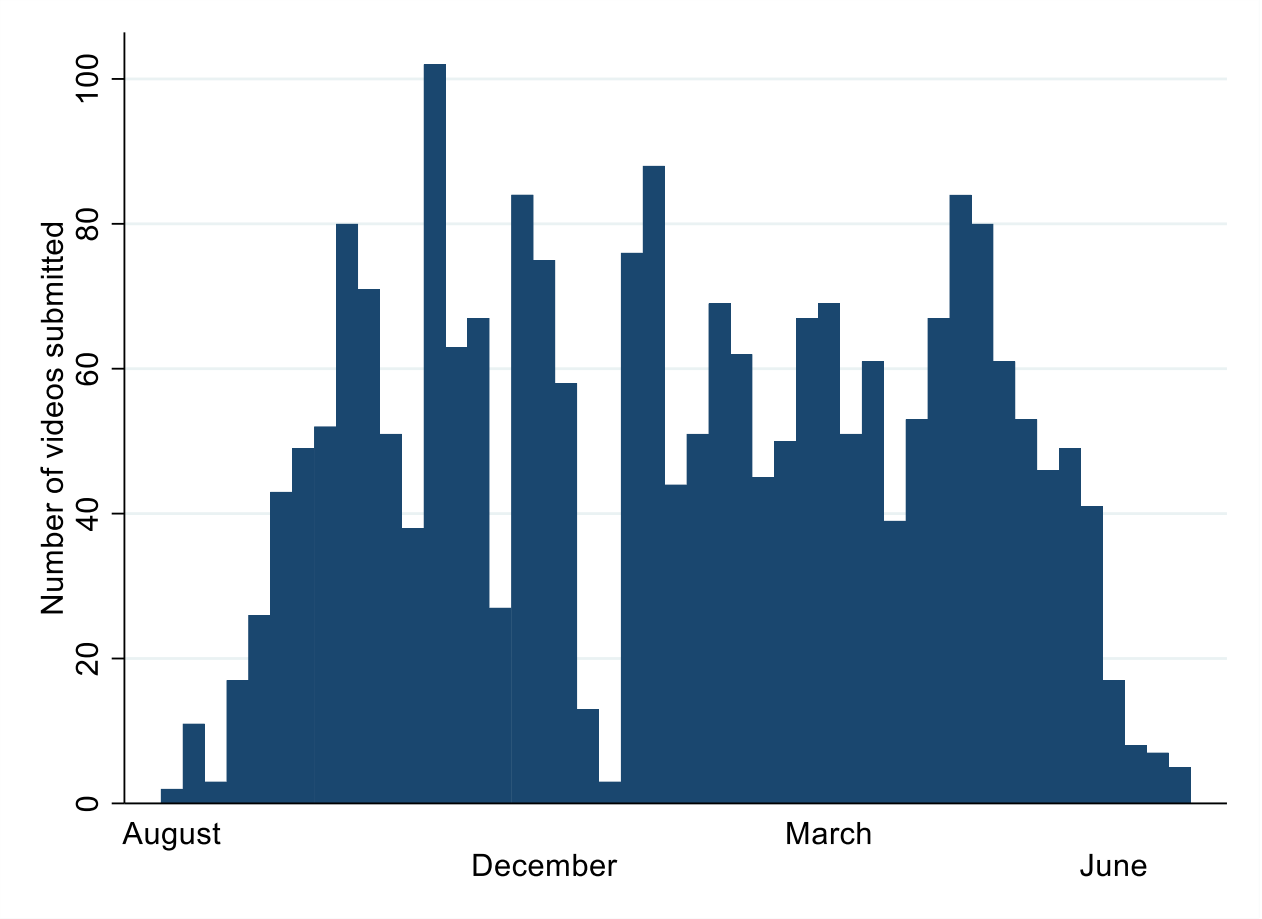


Figure A2. CLASS video submissions over time.

Note: Number of video submissions binned by week.

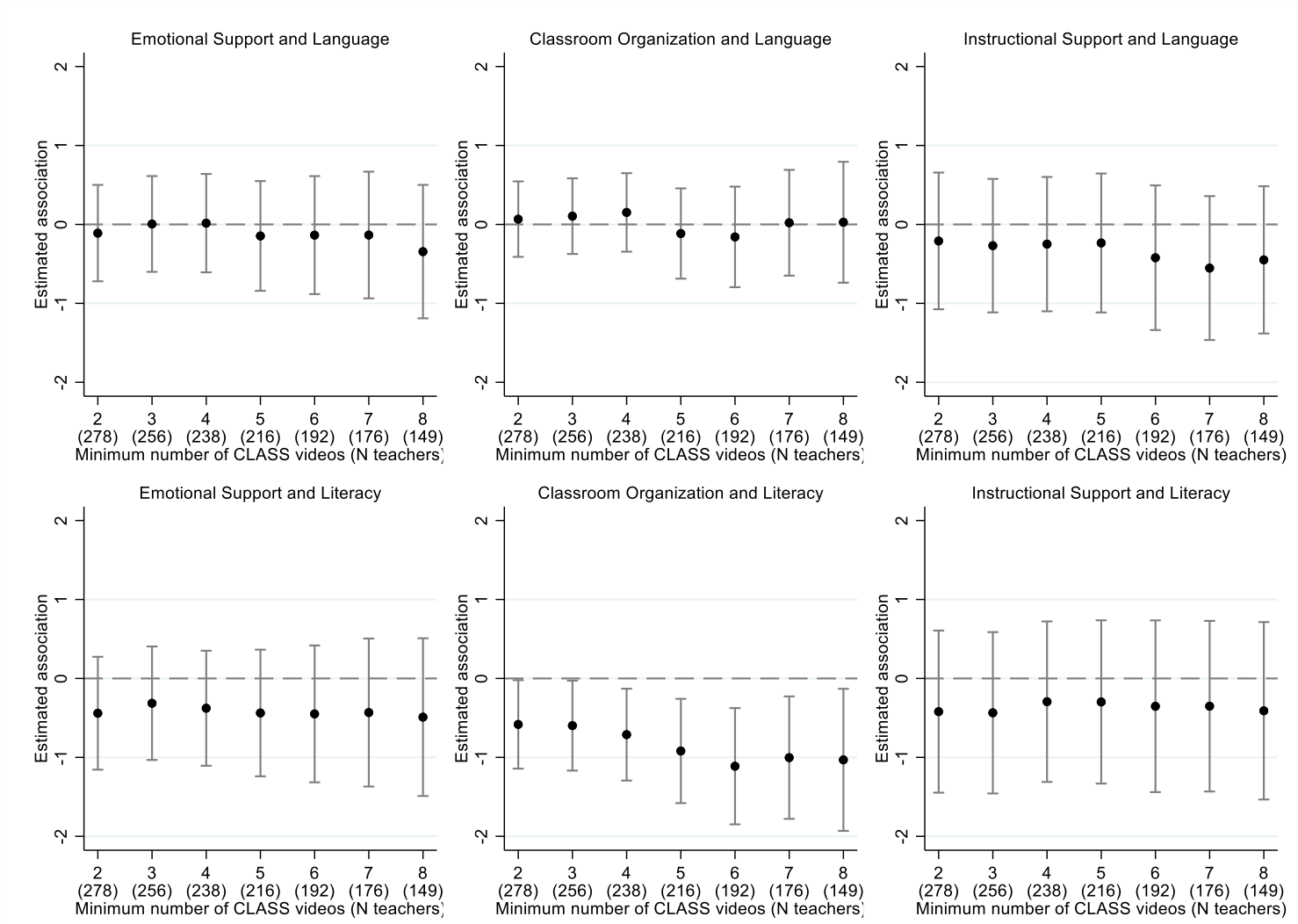


Figure A3. Estimated association between overall variation in quality and children's composite language and literacy scores, restricting the sample to teachers with minimum numbers of CLASS video submissions

Note: Figure shows point estimates and .95 confidence intervals. The leftmost bars correspond to the results shown in the main paper.

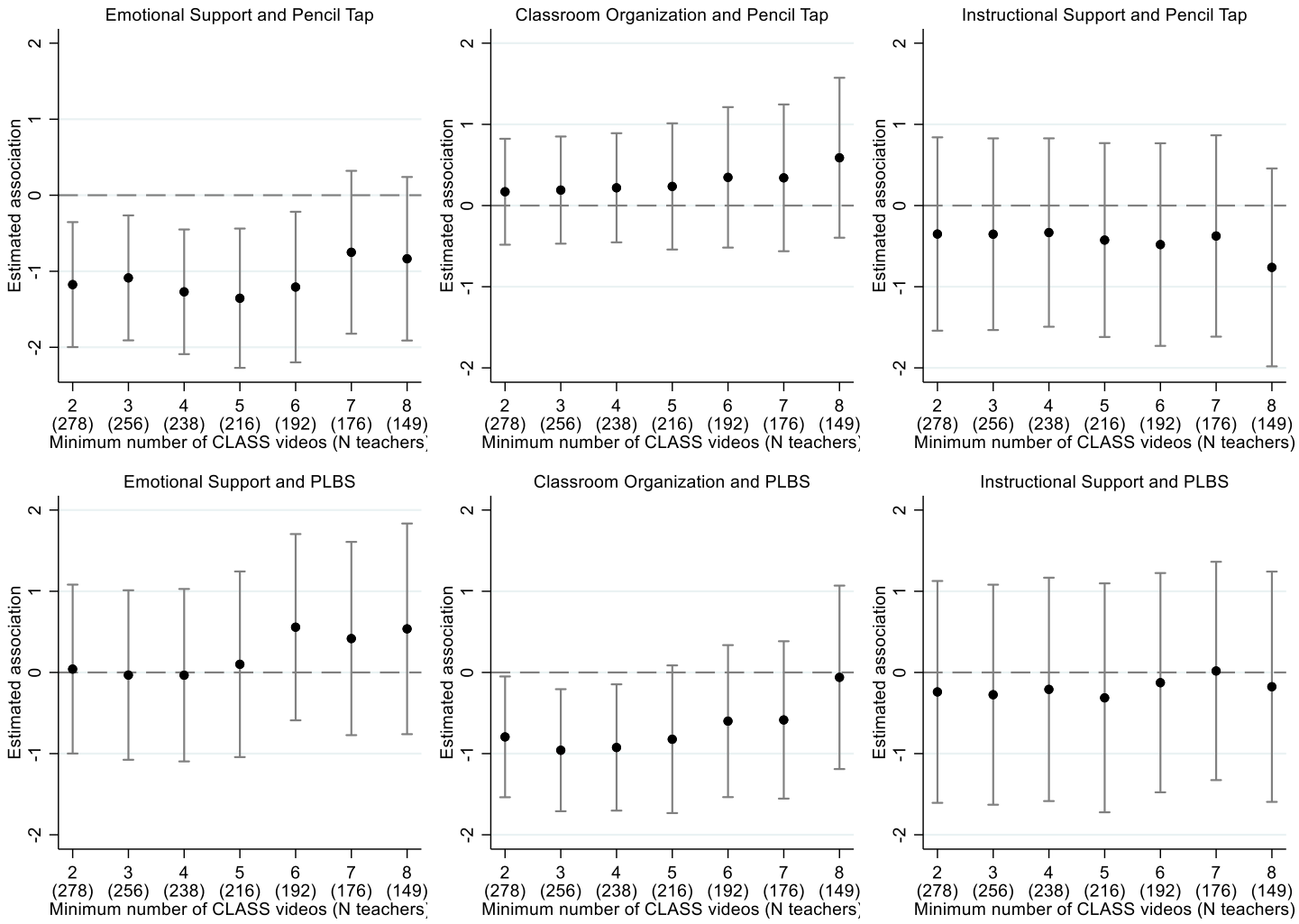


Figure A4. Estimated association between overall variation in quality and children's Pencil Tap and PLBS scores, restricting the sample to teachers with minimum numbers of CLASS video submissions

Note: Figure shows point estimates and .95 confidence intervals. The leftmost bars correspond to the results shown in the main paper.

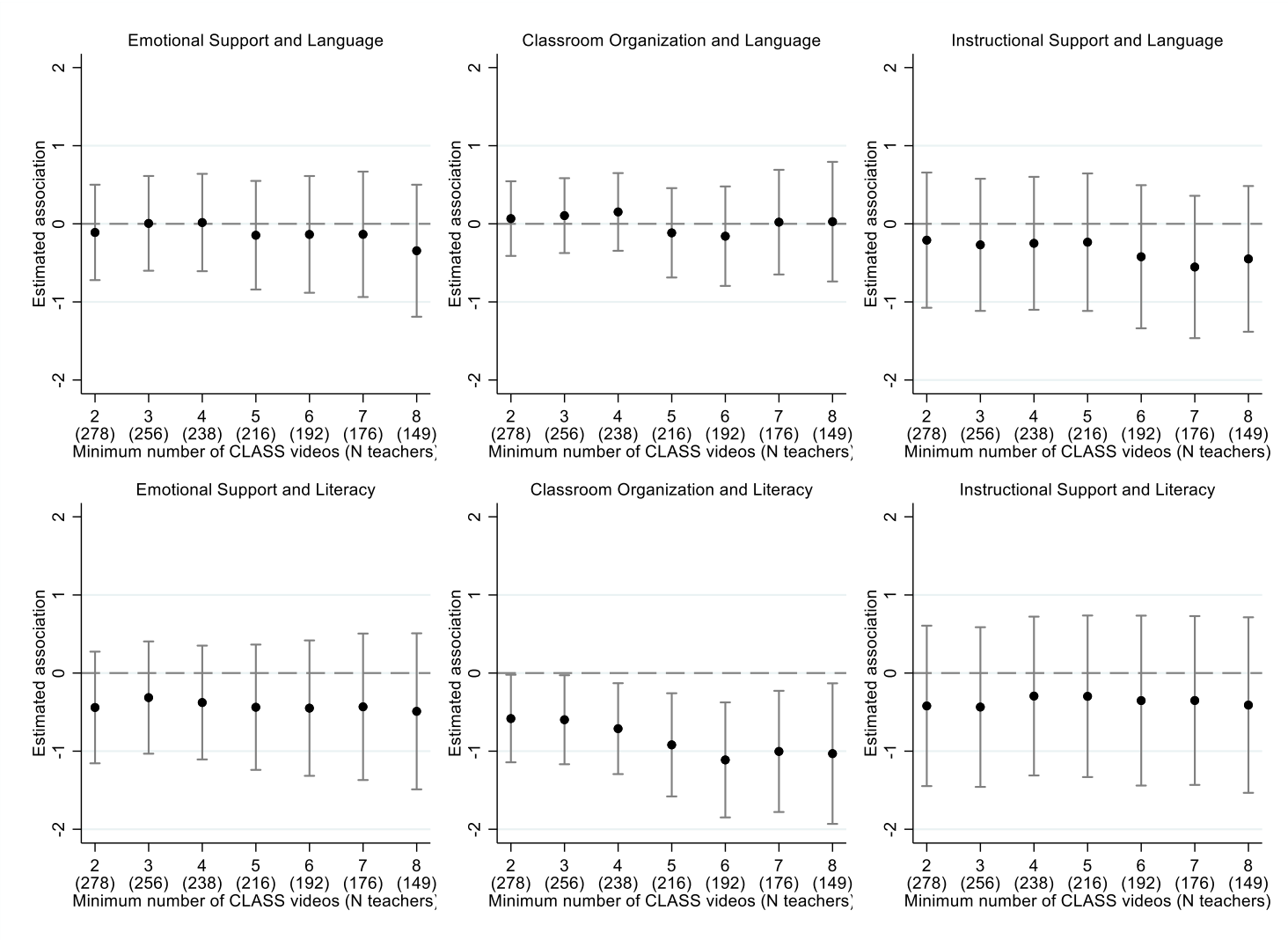


Figure A5. Estimated association between instability in quality and children's composite language and literacy scores, restricting the sample to teachers with minimum numbers of CLASS video submissions

Note: Figure shows point estimates and .95 confidence intervals. The leftmost bars correspond to the results shown in the main paper.

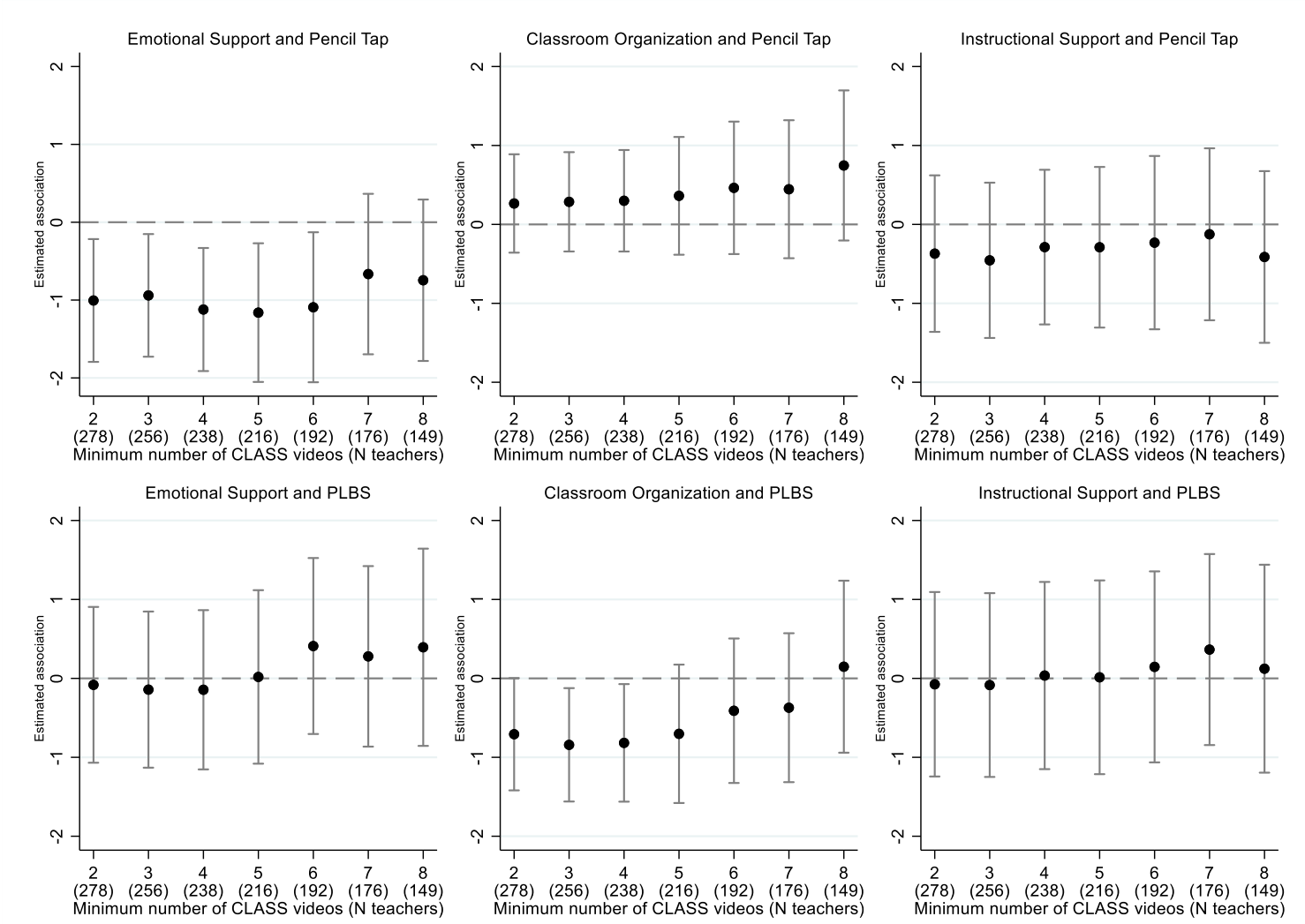


Figure A6. Estimated association between instability in quality and children's Pencil Tap and PLBS scores, restricting the sample to teachers with minimum numbers of CLASS video submissions

Note: Figure shows point estimates and .95 confidence intervals. The leftmost bars correspond to the results shown in the main paper.

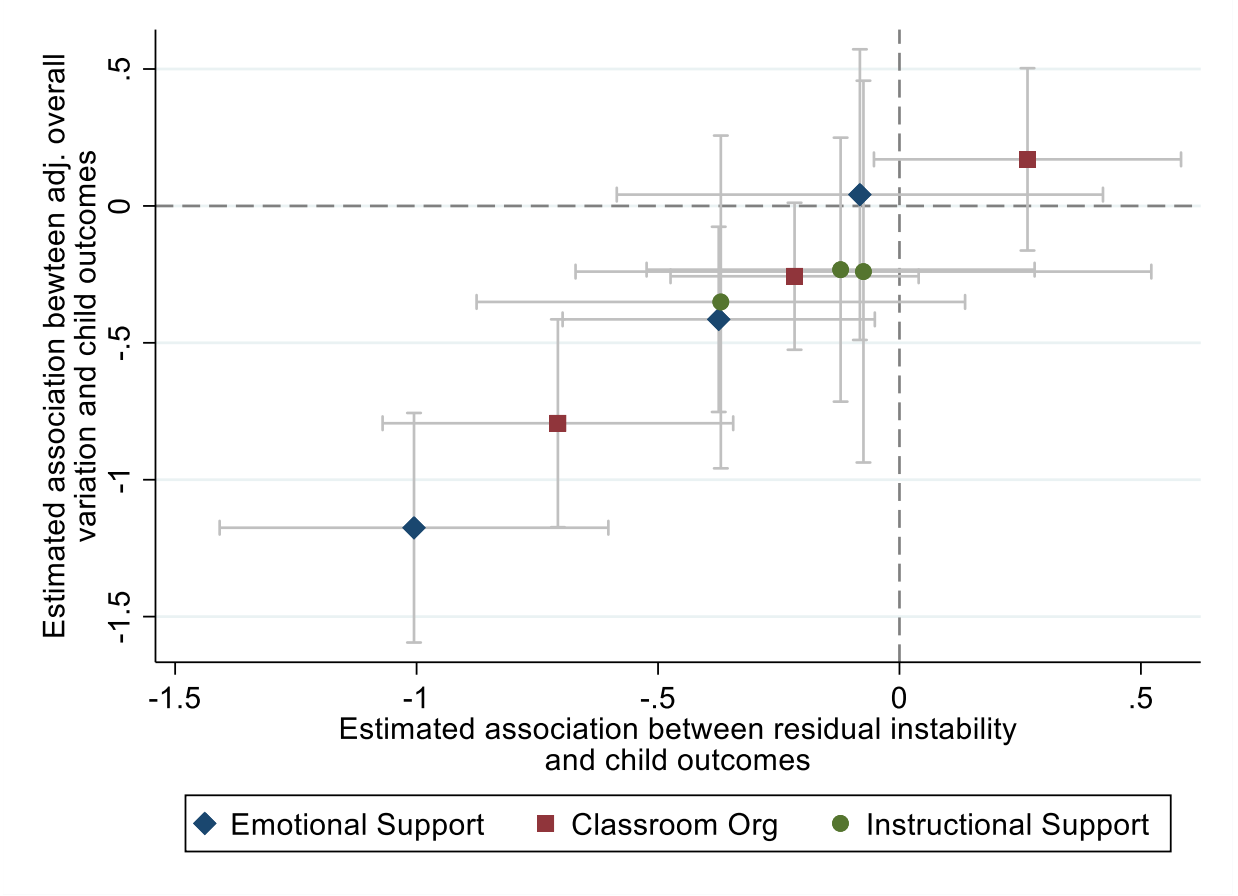


Figure A7. Estimated associations between variation in quality and child outcomes, using adjusted overall variation in quality and residual instability in quality

Note: Each point represents the estimated association between adjusted overall variation in quality/residual instability in quality for a unique pair of CLASS domain and child outcome measure. Bars indicate standard errors. These represent the associations presented in Table 3 (adjusted overall variation) and Table 5 (residual instability). Child outcomes include Pencil Tap, PLBS, and composite language and literacy scores.

Details of models examining associations between teacher-specific trends and instability in quality, and child outcomes

Associations between teacher-specific trends in quality and child outcomes

We used a straightforward regression-based approach to examine the association between teacher-specific trends in quality and child outcomes, Specifically, we estimated a two-level model of the following form:

Model A1:

$$\text{(Level 1)} \quad Y_{ij} = \beta_{0j} + \beta_1 \overrightarrow{X_{ij}} + \epsilon_{ij}$$

$$\text{(Level 2)} \quad \beta_{0j} = \pi_0 + \pi_1(10 * RETime_j) + \pi_2 AvgCLASS_j + \pi_3 SqrtNumCLASS_j + \pi_4 AvgCLASS_j * SqrtNumCLASS_j + \pi_5 \overrightarrow{T_j} + \pi_6 Treat_j + u_{0j}$$

where $Y_{ij}$is the outcome for child *i* in classroom with teacher *j*. $RETime_j$ is the Empirical Bayes estimate of the random time effect for teacher *j*. To clarify the interpretation of π_1, $RETime_j$ was multiplied by 10 such that it can be interpreted as the teacher-specific change in CLASS scores over (roughly) the full school year, beyond the average change of teachers in the same treatment condition. We also controlled for the average of teachers' observed CLASS scores ($AvgCLASS_j$), the square root of the number of video submissions ($SqrtNumCLASS_j$), and their interaction. We finally controlled for set of child-level covariates, X_{ij}, and a set of teacher- and school-level covariates, $\overrightarrow{T_j}$. Therefore, π_1 is the coefficient of interest and represents the association between teacher-specific trends in quality and child outcomes, above and beyond average quality over the year and other measured teacher characteristics.

Associations between instability in quality and child outcomes

Following our original model of overall variation, we used a similar approach to examine the association between instability in quality and child outcomes beyond teacher-specific trends. Specifically, we estimated a two-level model of the following form:

Model A2:

$$\text{(Level 1)} \quad Y_{ij} = \beta_{0j} + \beta_1 \overrightarrow{X_{ij}} + \epsilon_{ij}$$

$$\text{(Level 2)} \quad \beta_{0j} = \pi_0 + \pi_1 SDResid_j + \pi_2 AvgCLASS_j + \pi_3 SqrtNumCLASS_j + \pi_4 AvgCLASS_j * SqrtNumCLASS_j + \pi_5 \overrightarrow{T_j} + \pi_6 Treat_j + u_{0j}$$

where $Y_{ij}$is the outcome for child *i* in classroom with teacher *j*. $SDResid_j$ is the adjusted standard deviation of the residuals for teacher *j*. All other variables are defined as above. The parameter π_1 is the coefficient of interest and represents the association between instability in quality and child outcomes, above and beyond average quality over the year and other measures covariates.

Details of models examining associations between teacher participation in coaching and variation in quality

To estimate the association between teachers' participation in the coaching condition (i.e., assignment to the treatment condition) and overall variation in quality, we estimated a model of the following form for teacher *j*:

Model A3:

$$AdjSDCLASS_j = \beta_0 + \beta_1 Treat_j + \beta_2 \overrightarrow{T_j} + \epsilon_j$$

where $AdjSDCLASS_j$ is the adjusted measure of overall variation in CLASS scores for teacher *j*, and $\overrightarrow{T_j}$ is a vector of teacher- and school-level covariates including teacher age, experience, education, whether the teacher taught in a Head Start center, whether the teacher taught in a public school, classroom poverty, treatment status in the first phase of the study, whether the teacher was added to the study in the second phase, and site (city) indicators.

To estimate the association between teachers' participation in coaching condition (i.e., assignment to the treatment condition) and overall variation in quality, we estimated a model of the following form:

Model A4:

$$SDResid_j = \beta_0 + \beta_1 Treat_j + \beta_2 \overrightarrow{T_j} + \epsilon_j$$

Where $SDResid_j$ is the adjusted measure of instability in CLASS scores for teacher *j*, and $\overrightarrow{T_j}$ is the same vector of teacher- and school-level covariates described above.

Appendix B. Calculating adjusted measures of overall teacher-specific variation

Our first research aim focuses on examining the association between overall variation in classroom quality (i.e., teacher CLASS score variance) and child outcomes. However, teachers' estimates of overall variation are based on varying number of video submissions. To account for differences in precision across teachers due to varying number of video submissions, we calculate adjusted measures of teachers' overall CLASS score variation. This adjustment shrinks estimates of overall variation for teachers with fewer video submissions towards the estimated cross-teacher average of teacher-specific CLASS score variation. The adjustment was done using a three-step process outlined below, based on classic multilevel modeling techniques as described in, e.g., Raudenbush & Bryk (2002).

Step 1: Estimate standard error of the variance for each teacher *j*

We begin with estimates of overall CLASS score variation for each teacher *j* ($\widehat{s_j^2}$), calculated as the sample variance of CLASS scores from videos submitted by teacher *j*.

The standard error of each teacher's CLASS score variance was calculated by the following:

$$\widehat{SE}(\widehat{s_j^2}) = \sqrt{\frac{2(s_*^2)^2}{n_j - 1}}$$

where s_*^2 is average of the observed teacher-specific CLASS score variances, across all teachers, and n_j is the number of video observation for teacher *j*. We pool to ensure stability in estimating the standard errors; otherwise, a low estimate of s_j^2 will give a spuriously low standard error estimate as well. The above shows that the standard error of the CLASS score variance for a given teacher decreases with the number of video submissions from that teacher.

Step 2: Calculate shrinkage factor for each teacher *j*

The shrinkage factor for teacher *j* is given by the following:

$$\lambda_j = \frac{\tau^2}{\tau^2 + \widehat{SE}^2(\widehat{s_j^2})}$$

where $\widehat{SE}^2(\widehat{s_j^2})$ is the above estimated squared standard error of the variance for teacher *j*, and τ^2 is a method-of-moments estimate of the between-teacher variance in teacher-specific CLASS score variances. This estimate was calculated by the following:

$$\tau^2 = Var\left(\widehat{s_j^2}\right) - \frac{1}{J}\sum_{j=1}^{J} \widehat{SE}^2\left(\widehat{s_j^2}\right)$$

where $Var\left(\widehat{s_j^2}\right)$ is the variance of estimated teacher-specific CLASS score variances.

Our method-of-moments estimates of τ^2 for the between-teacher variance in the teacher-specific CLASS scores variances are as follows:

$\tau_{ES}^2 = 0.020$

$\tau_{CO}^2 = 0.046$

$\tau_{IS}^2 = 0.019$

Step 3: Calculate shrunken estimate of teacher-specific CLASS score variance

Finally, we applied the shrinkage factors to teachers' estimates of overall CLASS score variance. Specifically, the adjusted estimate of the teacher-specific CLASS score standard deviation for teacher *j* was calculated by the following:

$$\widehat{s_{EB,J}} = \sqrt{s_*^2 + \lambda_j\left(\widehat{s_j^2} - s_*^2\right)}$$

where s_j is the observed standard deviation of CLASS scores for teacher *j*, and s_*^2 is the average observed variance of CLASS scores, across all teachers.

The adjusted estimates of teacher-specific CLASS score standard deviations ($\widehat{s_{EB,J}}$) represents the key predictor of interest for the first research aim.

Appendix C. Sensitivity check for multiple hypothesis testing

We conduct a series of permutation tests to further test whether the results described in the text could reflect spurious correlations in the data rather than true underlying associations. In this Appendix, we describe the approach used to check the robustness of results regarding the first research question (associations between overall variation in quality and child outcomes). A similar exercise was used to test the robustness of results regarding the third research question (associations between residual instability in quality and child outcomes).

Step 1. For each of the three CLASS domains, we took the average of the estimated associations between overall variation in quality and scores on the Pencil Tap, PLBS, language composite, and literacy composite:

$$\hat{b}_{Avg,Domain} = \frac{\hat{b}_{PT,Domain}+\hat{b}_{PLBS,Domain}+\hat{b}_{Lang,Domain}+\hat{b}_{Lit,Domain}}{4} ,$$

where $\hat{b}_{PT,Domain}$, $\hat{b}_{PLBS,Domain}$, and $\hat{b}_{Lit,Domain}$ are the estimated associations between overall variation in the relevant CLASS domain and child Pencil Tap scores (based on estimating Model 1), child PLBS scores, child language scores, and child literacy scores. These averages are our test statistics; we want to test whether these average associations are larger than we would have seen due to random chance.

Step 2. For an overall test across all three measures of teacher quality, we calculated the average of the estimated average associations between overall variation in quality and child outcomes, across the three CLASS domains:

$$\hat{b}_{All} = \frac{\hat{b}_{Avg,ES} + \hat{b}_{Avg,CO} + \hat{b}_{Avg,IS}}{3}$$

where $\hat{b}_{Avg,Domain}$ are as described in Step 1.

Step 3. We then randomly permuted the three measures of variation in classroom quality across teachers. To account for the fact that measures of overall variation are correlated across the three CLASS domains, we permuted the three measures of overall variation in quality jointly. Therefore, each teacher was randomly assigned the values of overall variation in emotional support, classroom organization, and instructional support from another teacher. Values of all covariates and child outcomes are unchanged. This permutation approach preserves the structure of the data, in particular the correlation of students within classrooms and how the quality measures co-occur.

Step 4. We then re-estimated our primary model using the permuted values of overall variation in CLASS scores and the observed values of covariates and child outcomes. Using these new estimates of association, we then re-calculated the average of the associations between

variation in each CLASS domain and the three child outcomes, using the same approach as in Step 1, above. We also re-calculated the average of the estimated average associations between overall variation in quality and child outcomes across the three CLASS domains, using the same approach as in Step 2, above.

Step 5. We repeat Steps 3 and 4 over 1,000 iterations. We then compared the observed average association between variation in CLASS scores and child outcomes to the distribution of estimates from our simulated data. If our originally observed association is in the extreme tail of the distribution, we reject the null that there is no association between at least some of the variation in quality measures and at least some of the child outcomes.

This testing procedure is testing whether the classroom quality variation is associated with child outcomes *overall*. In particular, we find mild evidence that some child outcomes *are* associated with emotional support ($p \approx 0.072$ for a two-sided test) and classroom organization ($p \approx 0.076$, two-sided), but not for instructional support. When we test overall association across all 12 outcomes to protect against multiple testing we obtain $p \approx 0.020$, two-sided, indicating a real negative relationship between variation in teacher quality and child outcomes overall.

Our conditional version of this test, of variation beyond year trends, is less conclusive. We cannot completely rule out the possibility that year trends do explain all the relationship of classroom variation and child outcomes (the overall two-sided p-value is 0.056), although we do see some evidence that remaining variation in emotional support is connected to outcomes ($p \approx 0.073$). We note that given a belief that *increased* variation could not improve child outcomes, we could use a one-sided test and our p-values, making them land below the canonical 0.05 threshold.

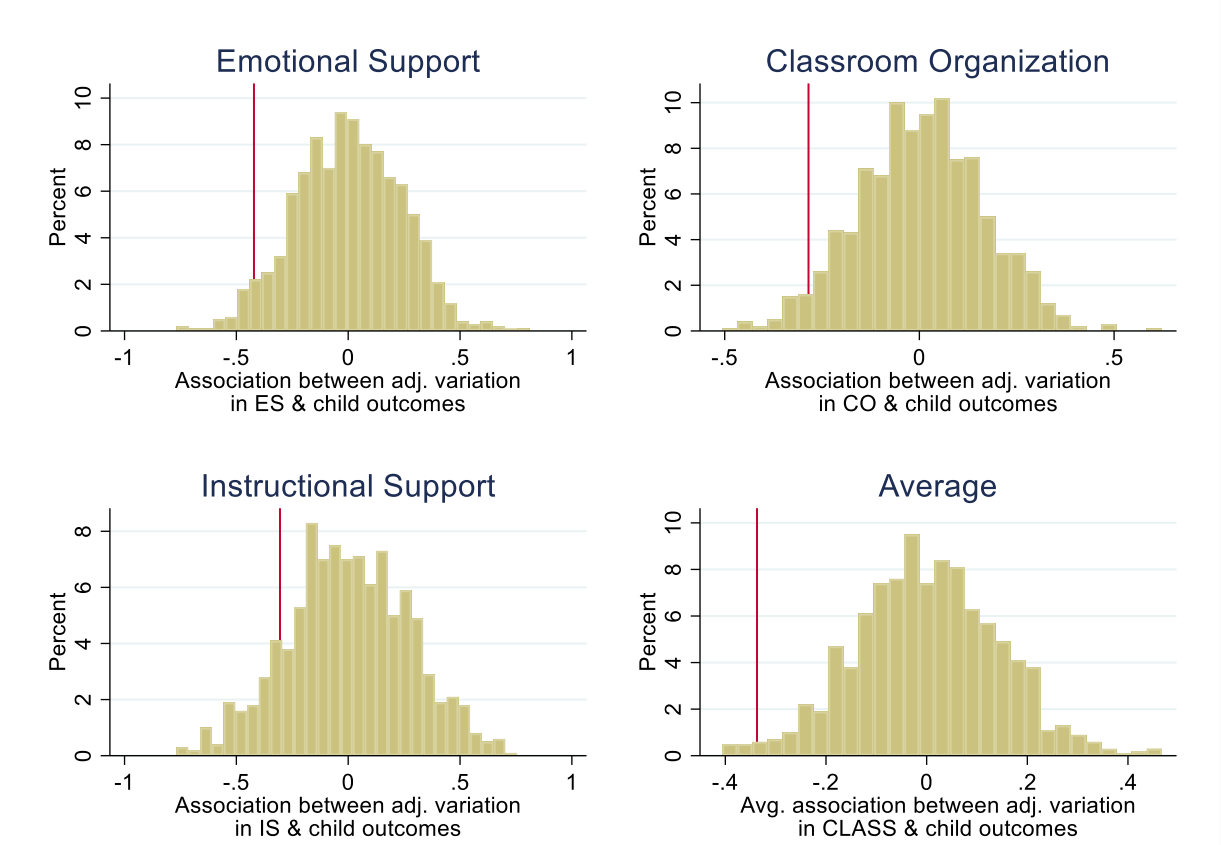


Figure C1. Results of simulation exercise examining association between overall variation in quality and child outcomes

Note: Vertical line indicates observed average of estimated associations between overall variation in the relevant CLASS domain and child outcomes (Pencil Tap, PLBS, and language and literacy composite scores). For emotional support, 7.2 percent of simulation results were more extreme (in either the left of right tail) than the observed average association. For classroom organization, 7.6 percent of simulation results were more extreme than the observed average association. For instructional support, 26.2 percent of simulation results were more extreme than the observed average association. For the average association across the three CLASS domains, 2.0 percent of simulation results were more extreme than the observed average association.

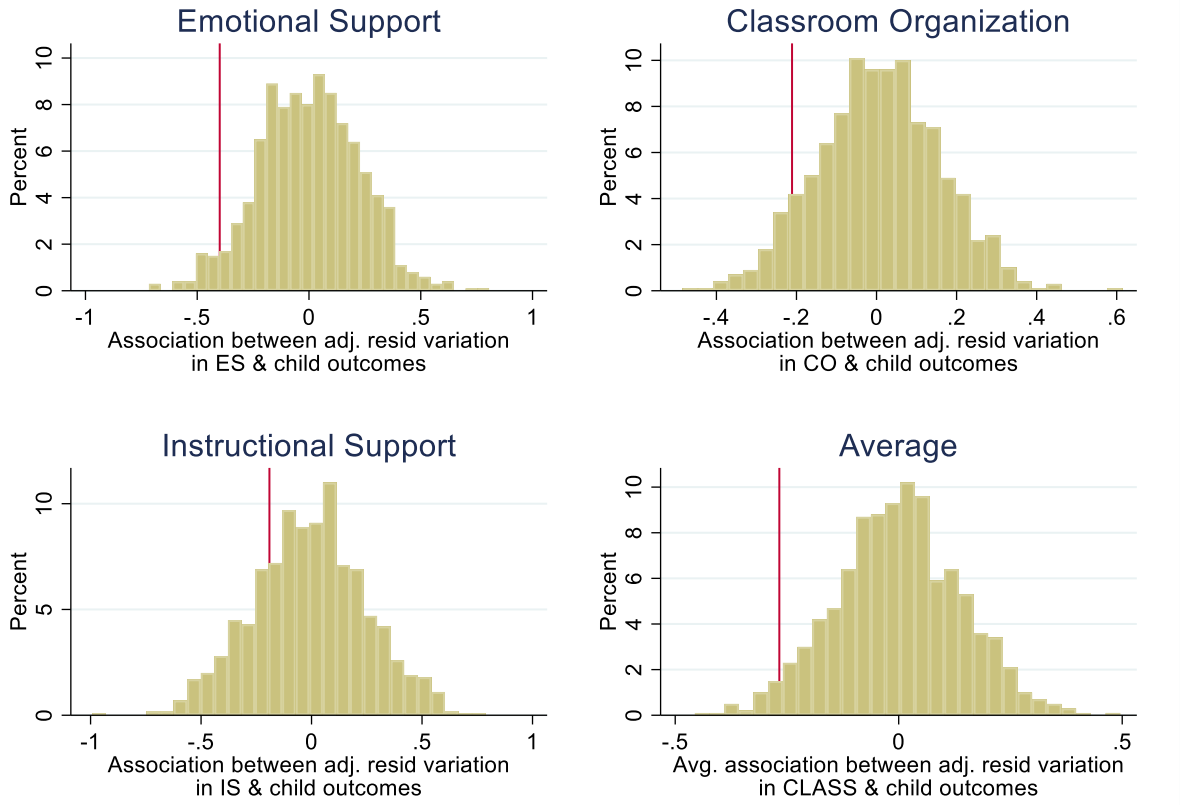


Figure C2. Results of simulation exercise examining association between residual instability in quality and child outcomes

Note: Vertical line indicates observed average of estimated associations between residual instability in the relevant CLASS domain and child outcomes (Pencil Tap, PLBS, and language and literacy composite scores). For emotional support, 7.3 percent of simulation results were more extreme (in either the left of right tail) than the observed average association. For classroom organization, 17.9 percent of simulation results were more extreme than the observed average association. For instructional support, 44.9 percent of simulation results were more extreme than the observed average association. For the average association across the three CLASS domains, 5.6 percent of simulation results were more extreme than the observed average association.