CHILD DEVELOPMENT

Child Development, xxxx 2021, Volume 00, Number 0, Pages 1–18

Examining the Effects of Changes in Classroom Quality on Within-Child Changes in Achievement and Behavioral Outcomes

Tyler W. Watts
Teachers College, Columbia University

Tutrang Nguyen
Mathematica

Robert C. Carr
Duke University

Lynne Vernon-Feagans
University of North Carolina at Chapel Hill

Clancy Blair
NYU School of Medicine

This study examines whether changes in classroom quality predict within-child changes in achievement and behavioral problems in elementary school (ages spanning approximately 6–11 years old). Drawing on data from a longitudinal study of children in predominantly low-income, nonurban communities (n = 1,078), we relied on child fixed effects modeling, which controlled for stable factors that could bias the effects of classroom quality. In general, we found that changes in classroom quality had small and statistically nonsignificant effects on achievement and behavior. However, we found that moving into a high-quality classroom, particularly those rated as high in *Classroom Organization*, had positive effects on achievement and behavior for children with significant exposure to poverty in early life.

Educational research has often touted the importance of high-quality classroom environments for promoting children's achievement and behavioral outcomes (e.g., Ansari & Pianta, 2018; Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008), especially for those growing up in disadvantaged communities (Hamre & Pianta, 2005). Indeed, developmental theory posits that classroom quality plays an important role in helping children reach their potential (e.g., Pianta & Hamre, 2009). Researchers have identified multiple proximal classroom processes that could support healthy cognitive and behavioral development, including teacher warmth and responsiveness, cognitive stimulation, classroom orderliness, and opportunities for learning (Mashburn et al., 2008).

Yet, evidence linking specific features of the classroom to child achievement and behavior has come primarily from correlational designs (e.g., Curby, Brock, & Hamre, 2013; Mashburn et al., 2008; Rimm-Kaufman, Baroody, Larsen, Curby, & Abry, 2015; Vernon-Feagans, Mokrova, Carr, Garrett-Peters, & Burchinal, 2019). These studies rely on control variables to account for nonrandom selection into classrooms, but such approaches are unlikely to account for all of the unobserved factors that lead some children to enroll in higher quality classrooms (see Rothstein, 2009). Thus, most correlational studies leave open questions regarding the causal effects of specific classroom processes on child outcomes, making it difficult to inform interventions that seek to improve features of children's educational environments.

This study attempted to further advance our understanding of the relation between classroom

This research was initially supported by a grant from the National Institute of Child Health and Human Development (NICHD), 1PO1HD39667 and 2PO1HD039667. Co-funding was provided by the National Institute of Drug Abuse, NIH Office of Minority Health, NIH-Office of the Director, National Center on Minority Health and Health Disparities, and the Office of Behavioral and Social Sciences Research. The current support comes from NICHD HD R01HD080786. We thank Drew Bailey, Peg Burchinal, Greg Duncan, and Bob Pianta for their helpful comments on previous drafts. We would like to express our gratitude to all of the families, children, and teachers who participated in this research and to the Family Life Project (FLP) research assistants for their hard work and dedication to the FLP. This study is part of the Family Life Project (https://flp.fpg.unc.edu/).

Correspondence concerning this article should be addressed to Tyler W. Watts, Teachers College, Columbia University, 525 W 120th St, New York, NY 10027. Electronic mail may be sent to tww2108@tc.columbia.edu.

 0009-3920/2021/xxxx-xxxx
DOI: 10.1111/cdev.13552

processes and child outcomes by employing within-child modeling—a strategy that has been recently used in other studies of child development to improve causal inference (see Maldonado-Carreño & Votruba-Drzal, 2011; Zachrisson & Dearing, 2015). Specifically, we tested the effects of changes in observed classroom quality on concurrent changes in child outcomes using child fixed effects models, which controlled for all stable characteristics of children and their environments that could bias observed associations between classroom processes and developmental outcomes. As we describe in the following section, we tested whether improvements in specific dimensions of classroom quality produced benefits on measures of academic achievement and behavioral adjustment collected throughout elementary school, and we also tested whether these benefits were largest for children from the most disadvantaged homes.

Classroom Quality

To measure classroom quality, researchers have often relied on the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008), an observational measure of classroom process quality that focuses on capturing teacher–child interactions. The CLASS is organized into three domains, each measuring unique elements of classroom quality (Pianta & Hamre, 2009). The first domain, called *Emotional Support*, is guided by work on attachment (e.g., Ainsworth, Blehar, Waters, & Wall, 1978) and self-determination theories (e.g., Ryan & Deci, 2000), and it assesses the quality of the classroom climate and the emotional regard between teachers and students. The *Classroom Organization* domain is motivated by research showing the importance of promoting students' self-regulatory abilities (e.g., Blair & Raver, 2015), and largely captures teachers' approaches to managing time, behavior, and attention in their classrooms. Finally, the *Instructional Support* domain is based on theoretical and empirical work that has attempted to identify optimal approaches for stimulating student cognitive growth (e.g., Davis & Miyake, 2004), and it measures the quality of teachers' instructional practices and scaffolding strategies for cognitive and language development.

A growing empirical literature has demonstrated positive associations between the domains measured by the CLASS and measures of student achievement and behavior (Allen et al., 2013; Curby et al., 2013; Mashburn et al., 2008; Rimm-Kaufman et al., 2015; Vernon-Feagans et al., 2019). Previous research has shown that children in more engaging and supportive classrooms show greater gains in reading, math, executive functioning, and social skills (Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Burchinal et al., 2016; Hamre, Hatfield, Pianta, & Faiza, 2014; Mashburn et al., 2008; Pianta, Belsky, et al., 2008). Teachers who establish well-run, organized classrooms help to prevent behavior problems and maximize opportunities to learn (Rimm-Kaufman, Curby, Grimm, Nathanson, & Brock, 2009). These teaching practices have also been shown to help students reduce issues related to behavioral and emotional dysregulation, including the types of externalizing behaviors that can lead to poor long-term relationships with teachers and peers (Raver et al., 2009). Moreover, teachers who emphasize conceptual understanding, provide feedback, and engage children in conversations during instruction have been found to promote children's gains in literacy, language, and math outcomes (Burchinal et al., 2010; Mashburn et al., 2008).

However, some correlational work linking CLASS domains to child outcomes has also reported mixed or null findings (Guerrero-Rosada et al., 2021; Weiland, Ulvestad, Sachs, & Yoshikawa, 2013). Indeed, a recent meta-analysis of studies linking dimensions of the CLASS to child outcomes during preschool reported that average effects were often small (i.e., < 0.10; Perlman et al., 2016). Furthermore, the meta-analysis noted apparent signs of bias in the correlational literature largely due to confounding (i.e., inadequate controls) and sample selection. However, several studies also imply that average classroom quality effects may mask important sources of heterogeneity. Long-standing developmental theory (e.g., McCartney & Berry, 2009; Ramey & Ramey, 1998; Scarr & McCartney, 1983) and previous empirical work (Berry et al., 2016; Dearing, McCartney, & Taylor, 2009; Peisner-Feinberg et al., 2003) suggest that classroom and child-care experiences may be especially important for children from disadvantaged communities. In particular, Dearing et al. (2009) found evidence that high-quality care environments during early childhood could protect against the deleterious effects of high-risk home environments. Similarly, Berry et al. (2016) found that more time spent in formal child care buffered against the negative effects of chaotic home environments on children's cognitive and socioemotional development. These compensatory effects could arise if well-organized classrooms shaped by daily routines offset the negative influences of highly chaotic and

unpredictable home contexts on children's academic and behavioral functioning (Raver et al., 2009). Similarly, emotionally supportive classrooms, where teachers exhibit high degrees of warmth and emotional understanding, could be especially supportive of the behavioral regulation of children from homes taxed by the stress of poverty (see discussion of "prosocial" classrooms in Jennings & Greenberg, 2009).

Recent work on classroom quality has also found evidence of quality thresholds, suggesting that the practices captured by the CLASS differentially affect child outcomes above or below certain levels of quality (Broekhuizen, Mokrova, Burchinal, & Garrett-Peters, 2016; Burchinal et al., 2010, 2016; Hatfield, Burchinal, Pianta, & Sideris, 2016; Vernon-Feagans et al., 2019; Weiland et al., 2013). These threshold studies have found that classroom quality must reach a particular level before it can affect child outcomes. Using data from a multistate study of preschool programs, Burchinal et al. (2010) found stronger relations between *Instructional Quality* and academic achievement in higher quality preschool classrooms than in lower quality classes (though classes scoring as low as 3.25 on the 1–7 scale of *Instructional Quality* were still considered "higher quality" in this sample). A similar pattern was observed for the relation between *Emotional Support* and reductions in behavioral problems.

Finally, classroom quality may have stronger effects on child development at different ages. However, although multiple studies have tracked children's developmental outcomes and classroom quality longitudinally (e.g., Ansari & Pianta, 2018; Curby, Rimm-Kaufman, & Cameron Ponitz, 2009; Vernon-Feagans et al., 2019), little research has examined whether quality effects vary by age. Theoretical frameworks derived from psychology and economics converge on the idea that children derive greater benefit from developmentally promotive inputs experienced earlier in childhood compared to later in childhood (see Heckman, 2006; Ramey & Ramey, 1998). However, the evidence is mixed with regard to the timing of quality caregiving and educational inputs (e.g., Li, Farkas, Duncan, Burchinal, & Vandell, 2013; Rea & Burton, 2020). Given this mixed evidence, as well as a lack of previous research on classroom quality effects during elementary school, we conducted exploratory analyses to examine whether the effect of classroom quality differed across the elementary school grades considered in this study.

Classroom Quality Effects on Child Outcomes and Causal Identification

Although the studies reviewed above report compelling theoretical links between specific dimensions of classroom quality and student outcomes, lingering concerns over bias (see Perlman et al., 2016) and causal identification hamper our ability to draw strong theoretical and policy conclusions. Children are not randomly sorted into classrooms, and studies that control for an earlier measure of achievement still depend on the assumption that other unobserved factors do not bias associations between classroom quality and child outcomes. Several evaluations using experimental designs have shown that programs that boost classroom quality scores often produce concomitant boosts in child outcomes (Araujo, Carneiro, Cruz-Aguayo, & Schady, 2016; McCormick, Cappella, O'Connor, & McClowry, 2015; Pianta et al., 2017). However, some studies have reported disappointing results even when positive effects on CLASS scores were observed (see Kraft, Blazar, & Hogan, 2018; Yoshikawa et al., 2015). Unfortunately, intervention evaluations carry limited potential to generate theoretically informative estimates of the effect of specific dimensions of classroom quality on child development. Randomized control trials can only evaluate the specific intervention program in question, and consequently, cannot easily isolate whether the intervention affected child outcomes due to changes in classroom quality, or other factors associated with the program model (see McCormick et al., 2015).

This study attempted to isolate the unique effect of classroom quality on measures of child achievement and behavioral problems by employing a child fixed effects approach. This approach tests whether within-child changes in classroom quality between kindergarten and fifth grade predict within-child gains in achievement and behavioral outcomes during the same period. In these models, every child serves as their own control, effectively eliminating all stable sources of confounding variation (e.g., home environment or motivation) that could bias associations between measures of classroom quality and developmental outcomes.

Similar fixed effects models have been employed in other recent studies in developmental psychology to examine the effects of early child care on developmental outcomes (e.g., Zachrisson & Dearing, 2015), and economists have long used such approaches to try and examine the effects of educational inputs on student achievement (e.g., Boyd,

Lankford, Loeb, Rockoff, & Wyckoff, 2008; Rivkin, Hanushek, & Kain, 2005). Maldonado-Carreño and Votruba-Drzal (2011) used a similar within-child modeling framework to examine associations between child outcomes and teachers' ratings of their own relationships with students using the NICHD Study of Early Childcare and Youth Development. Relying on hierarchical linear modeling, they found only limited evidence that teacher–child relationships affected within-child growth in achievement, though effects on teacher-reported behavioral problems were larger. These findings reflect the conclusions of recent methodological studies showing that empirical approaches that control for stable variation between children produce more conservative estimates due to their bias-reducing capabilities (see Berry & Willoughby, 2017), and these estimates may better approximate the effects of interventions (Bailey, Duncan, Watts, Clements, & Sarama, 2018).

Current Study

We leveraged data from the Family Life Project (FLP), a longitudinal birth cohort study of children in nonurban areas of the United States The FLP contains measures of classroom quality and direct assessments of child functioning collected at multiple time points between kindergarten and fifth grade, allowing us to examine the effect of changes in classroom quality across the entire elementary school period. With these data, we examined several hypotheses that have been central to studies documenting relations between classroom quality and children's outcomes. First, building on work suggesting that high-quality childcare and classroom environments may protect against the negative effects of high-risk home environments (e.g., Dearing et al., 2009), we tested whether domains of classroom quality produced larger effects for children who have been exposed to substantial environmental disadvantage. Second, following the work on quality thresholds (Burchinal et al., 2010, 2016) we examined whether classroom quality effects differ across the quality distribution. Thus, we tested whether quality effects were consistent throughout the distribution of CLASS scores, or whether large changes in quality were required to access benefits.

This study builds upon previous work with the FLP sample that used regression-control approaches to examine the accumulation of high-quality educational experiences over multiple years. In particular, Broekhuizen et al. (2016) reported that CLASS scores, measured across preschool and kindergarten, predicted improvements in behavioral functioning at the end of kindergarten. Additionally, Vernon-Feagans et al. (2019) found evidence of quality thresholds, as they reported that additional years spent in high-quality classrooms from kindergarten through third grade predicted literacy achievement, especially for students who started school with lower reading skills. Whereas these studies relied on measures of cumulative classroom quality assessed over several years, this study relies exclusively on within-child variation, and provides estimates of the effect of within-child changes in classroom quality on corresponding changes in child functioning across multiple grades.

It should be noted that we approached our key models as largely confirmatory in nature. Although we did not preregister these analyses, we expected to see positive associations between CLASS scores and child outcomes using the child fixed effects model, and we also expected those associations to be larger for the most disadvantaged students. However, given the inconsistent findings from intervention studies (e.g., Yoshikawa et al., 2015), we expected the coefficients produced from our child fixed effects models to be smaller than those reported in other correlational work. Given the previous work in this area on quality thresholds (e.g., Burchinal et al. 2010, 2016), we also tested whether returns were largest at the top of the distribution of CLASS scores, though we did not have strong hypotheses regarding threshold effects. Furthermore, we tested if CLASS effects differed across grade levels, and we regarded those tests to be largely exploratory.

Method

Data

The FLP is a study of children and families living in two of the four major geographical areas of the United States with especially high rural poverty rates (Vernon-Feagans, Cox, Key, & Investigators., 2013). Specifically, three counties in Eastern North Carolina (NC) and three counties in Central Pennsylvania were selected to be indicative of the Black South and Appalachia, respectively. The FLP adopted an epidemiological design to recruit a sample representative of the population of children whose families resided in one of the six counties at the time of the child's birth (initial sampling occurred between the fall of 2003 and the fall of 2004). Low-income families were oversampled in

both states and African American families were oversampled in NC. Of 1,571 families selected for participation, 1,292 (82%) completed a home visit for data collection at age 2 months, at which point they were formally enrolled in the study. Study children were followed through elementary school. The current analysis utilized measures collected during early childhood home visits (at 6, 15, and 24 months) and subsequent school visits: kindergarten, Grade 1, Grade 2, Grade 3, and Grade 5 (Grade 4 information was not collected due to budgetary constraints). In some models, we also relied on cognitive and behavioral data collected during preschool as control variables. See Vernon-Feagans et al. (2013) for detailed information regarding the study design and sampling plan.

As we detail further in the following section, our current analytic sample (n = 1,078) was evenly split between boys and girls, and approximately 43% identified as Black and 55% as White. The average income-to-needs ratio during early childhood was 1.81. At the first-grade wave, the average age at the child assessment was 5.98 years old (SD = .30), and at the fifth-grade wave, the average age was 11.16 years old (SD = .37).

Measures

Classroom Quality

Observational assessments of classroom quality were collected using the CLASS (Pianta, La Paro, et al., 2008), which was administered at five timepoints during the school-year waves of data collection: kindergarten, first grade, second grade, third grade, and fifth grade. Observers for the FLP study participated in a 2-day training program conducted by a certified CLASS trainer. At the end of the 2-day training program, they completed a reliability test that involved each observer individually coding five 20-min video segments. The observer codes were then compared against the master codes provided by the CLASS trainers, and in order to pass the reliability test, observers had to reach an agreement within 1 point of each code on 80% of all codes across the five segments. Before each respective wave of data collection, observers recertified their training (see details in Vernon-Feagans et al., 2019).

Trained observers were tasked with observing each classroom for two 30-min cycles during reading instruction, and we averaged these two cycles together to create one observational record for each classroom. During the four waves between kindergarten and Grade 3, the majority of classroom observations occurred during the late fall. During the fifth-grade wave, observations occurred during the spring. The CLASS was scored in ten dimensions and organized into three domains: (a) *Emotional Support* (positive climate, negative climate [reversed], teacher sensitivity, and regard for student perspectives); (b) *Classroom Organization* (behavior management, productivity, and instructional learning formats); and (c) *Instructional Support* (concept development, quality of feedback, and language modeling). For each domain, classrooms were rated on a scale ranging from "1" to "7." Scores of "1" to "2" represent low quality, scores of "3" to "5" represent medium quality, and scores of "6" to "7" represent high quality. For the FLP sample, previous studies have reported high item-level consistency within each domain (α = .78–.84; Vernon-Feagans et al., 2019).

To examine the potential effects of thresholds in classroom quality, we tested models that relied on splitting CLASS scores into quartiles. Because the CLASS has been validated for multiple grades of elementary school (Pianta, La Paro, et al., 2008), we split the sample into quartiles across the entire distribution of CLASS scores collected during the elementary school years (i.e., kindergarten through Grade 5). For *Emotional Support* and *Classroom Organization*, the highest quartile groups included classes scoring at or above a "6," and for *Instructional Support*, the highest quartile group scored at "4" or above. Table S1 presents complete descriptive information for the each domain's quartile groups.

Academic Achievement

Academic achievement was measured using the Woodcock–Johnson III (WJ-III), a commonly used assessment battery of cognitive skills and academic ability for school-aged children (Woodcock, McGrew, & Mather, 2001). The WJ-III was administered via a one-on-one assessment with a trained examiner during each wave of school data collection. The WJ-III was scaled using a Rasch scoring method, and the "W scores" were derived as a direct transformation of the Rasch model (see Woodcock et al., 2001).

The FLP administered different subtests of the WJ-III during different waves. In each wave between preschool and Grade 5, the *Letter-Word Identification* subtest and the *Applied Problems* subtest were both administered. The *Letter-Word Identification* subtest asked students to name letters and read words, and the *Applied Problems* subtest

measured mathematical procedural knowledge. Beginning in kindergarten, the *Picture Vocabulary* subtest, a measure of receptive vocabulary, was also administered. Finally, the *Passage Comprehension* subtest, a measure of reading comprehension, was administered beginning in first grade. In third grade, a planned missingness procedure was used due to budgetary constraints (see Vernon-Feagans et al., 2019), and approximately half of the children in the current sample were randomly selected to receive the *Applied Problems* and *Picture Vocabulary* subtests (n = 429), whereas the other half took the *Letter-Word Identification* and *Passage Comprehension* subtests (n = 451). In fifth grade, the study returned to administering all of the aforementioned subtests to the entire sample.

Because we were interested in the effect of classroom quality across domains of academic achievement, and because we observed strong correlations among the WJ-III subtests across waves (r = .71–.91; see Supporting Information), we generated a composite measure of academic achievement. This achievement index was generated by first averaging together the W-scores for all nonmissing reading tests at each wave (i.e., *Letter-Word Identification, Picture Vocabulary,* and *Passage Comprehension*) to calculate a composite reading score. We then averaged this reading measure with the *Applied Problems* subtest to calculate an achievement composite variable that weighted math and reading scores equally. In the Supporting Information, we detail the results from models that examined CLASS effects on each WJ-III subtest individually.

Behavioral Problems

At each wave between preschool and Grade 5, classroom teachers rated children's behavioral difficulties using the Strengths and Difficulties Questionnaire (SDQ; Goodman, 2001). The SDQ items covered 25 attributes that were organized into five domains of child behavioral functioning (the current analysis excludes the *Prosocial Scale*). The *Emotional Symptoms* domain described problems with internalizing, worrying, and negative emotions (e.g., "often seems worried"). The *Conduct Problems* domain captured issues with externalizing behavioral problems and anti-social behavior (e.g., "often loses temper"). The *Hyperactivity Scale* measured attention problems and impulsivity (e.g., "restless, overactive, cannot stay still for long"), and the *Peer Problems* domain measured social difficulties (e.g., "picked on or bullied by other children").

The four domains indicating behavioral difficulties were combined into the "Total Difficulties" score for a general measure of behavioral problems. The Total Difficulties score has been shown to strongly correlate with other measures of behavioral problems, such as the Total Problem Behaviors score from the Child Behavioral Checklist (r = .87; Goodman & Scott, 1999), and a recent review of the psychometric properties of the SDQ reported strong internal consistency for the Total Difficulties score across studies (α = .82; see Kersten et al., 2016). As with the achievement measures, we also report results for the subscores included in the Total Difficulties score in Supporting Information.

Early Risk

We measured early socioeconomic risk using a cumulative risk factor developed by Burchinal and Willoughby (2013) for the FLP sample. The cumulative risk factor was computed using factor analyses of seven indicators: mother's education, family income-to-needs-ratio, work hours, job prestige, household density, neighborhood safety, and whether the mother is partnered. These variables were standardized and summed at the age 6-, 15-, and 24-month time-points. We averaged the risk score across these waves to generate one measure of early environmental risk, and we considered children scoring in the top quartile as our "high-risk" group. Children scoring below the 25th percentile comprised the "lower risk" group. Below, we also detail the results from models that used alternative definitions of early environmental risk that tested the sensitivity of our results to the risk index created for the FLP sample.

Analysis

The full FLP sample included 1,292 children at the 2-month interview. We first limited the sample to children who had a nonmissing early cumulative risk score, as well as at least one valid WJ-III or SDQ observation in elementary school (n = 1,108 children; 86% of the original sample). For the final analytic sample, we then included children who had at least one concurrent classroom observation measure corresponding to a WJ-III measure of achievement (n = 1,070; 83% of the full sample), or one concurrent classroom observation measure corresponding to an SDQ rating of behavioral problems (n = 1,074).

Across the "achievement" and "behavioral" samples, a total of 1,078 children were included. Across

the five waves included in our analysis, children in the achievement models were observed an average of 4.31 times, and children in the behavioral models were observed an average of 3.99 times. As expected, children tended to have more observations in the earlier waves of the study due to study attrition. In kindergarten, we observed 1,016 children (79% of the full FLP sample) in our analytic sample, and by fifth grade, we observed 748 children (58% of the full FLP sample). In the Appendix S1, we detail the results from models that used multiple imputation to impute any missing CLASS or outcome data at each wave, thus allowing us to observe all 1,078 children at each of the five measurement points (see Tables S2 and S3).

Because the sampling procedures used for the FLP study were not tied to schools (i.e., students were recruited at birth from local hospitals), we observed relatively little classroom-level clustering. On average, we observed 1.93 children per class, and 56% of the classrooms had only one study child per class. In results available upon request, we examined models that adjusted for classroom-level clustering, and the results were nearly identical to our main estimates.

Child Fixed Effects Models

We used child fixed effects models (see Imai & Kim, 2019) to estimate the effect of within-child changes in classroom quality on within-child changes in achievement and behavior:

$$\text{Outcome}_{ijt} = a_1 + \beta_1 \text{CLASS}_{ijt} + \pi\text{Child}_i + e_{ijt}, \quad (1)$$

where Outcome_{ijt} represents the achievement or behavior score for child i (n = 1,078) in classroom j (n = 2,450) at time t (either K, G1, G2, G3, or G5). In these models, children could be observed up to five times between kindergarten and Grade 5, with the CLASS observations (CLASS_{ijt}) and behavioral and cognitive assessments (Outcome_{ijt}) occurring during the same school year. Importantly, πChild_i captures the set of child fixed effects, which includes a unique intercept for each child. In essence, this model could be estimated by differencing each variable from every child's "grand mean" across all five study waves.

By including fixed effects, every child is given a unique intercept that controls for person-specific stable variation that could bias estimates. This means that stable characteristics about children, their families, or their environments are effectively "differenced out." Thus, with each child serving as their own control, all between-child differences in classroom quality and developmental outcomes are removed from the estimation, and β_1 measures the within-child effect of single-year changes in classroom quality on changes in achievement or behavior (see Figure S1 for a graphical representation of the model). It should be noted that Equation 1 also assumes that β_1 is stable across waves, though we later explore this assumption by including interactions between wave and classroom quality.

In Equation 1, β_1 could contain bias if certain factors are left unaddressed. First, time-varying factors (e.g., changes in child motivation) could bias estimates to the extent that these factors cause changes in both classroom quality and the observed outcomes. Second, because we rely on concurrent measures of child outcomes and classroom quality, our modeling approach assumes that the direction of the effect has been correctly specified (i.e., achievement should not cause concurrent classroom quality). Relatedly, if cross-lagged paths between classroom quality and child outcomes exist (e.g., classroom quality at Time 1 causes variation in achievement at Time 2, or achievement at Time 1 causes variation in classroom quality at Time 2), such effects could bias estimates if they were left unaccounted for. To address these potential sources of bias, we also include lagged measures of classroom quality, child achievement, and behavior:

$$\begin{aligned}\text{Outcome}_{ijt} = {} & a_1 + \beta_1 \text{CLASS}_{ijt} + \beta_2 \text{Ach}_{ijt-1} + \beta_3 \text{Beh}_{ijt-1} \\ & + \beta_4 \text{CLASS}_{ijt-1} + \beta_5 \text{Age}_{ijt} + \pi\text{Child}_i \\ & + \Omega\text{GradeLevel}_t + e_{ijt},\end{aligned} \quad (2)$$

where $\text{Ach}_{\text{ijt}-1}$ and $\text{Beh}_{\text{ijt}-1}$ are lagged measures of achievement and behavior, respectively, and $\text{CLASS}_{\text{ijt}-1}$ is a lagged measure of classroom quality. With this model, β_1 represents the effect of single year *changes* in classroom quality on *changes* in child outcomes, controlling for stable child characteristics. Here, we also include a measure of the child's age at the time of the outcome assessment ($\beta_5\text{Age}_{\text{ijt}}$), and a set of dummy variables for study wave ($\Omega\text{GradeLevel}_t$) to account for possible differences between grades that could affect results.

In all models, we adjusted standard errors for student-level clustering (i.e., the repeated observations over time for each student). Outcome variables were within-grade standardized across the entire sample. To allow for easy comparisons with other studies reporting effects of the CLASS measure, our main models included the CLASS

domains in raw units (i.e., varying from 1–7), but we also report standardized effects in Supporting Information.

As we describe in the following section, we present results that extended Equation 2 in several theoretically motivated ways. First, we present results from models that split the sample between the high- and lower risk groups to illustrate group differences based on environmental risk, and we used chi-square tests to test if model parameters were statistically significantly different between high- and lower risk children. We then report results for models that included an interaction between risk and classroom quality to test the magnitude of the difference in the classroom quality effect for the high-risk group. Second, we present estimates that included the three domains of the CLASS modeled simultaneously to estimate the unique effects of changes in each respective domain. Third, to test the possibility of quality threshold effects, we contrast estimates generated from continuous measures of CLASS scores (i.e., these models assume a linear return to changes in classroom quality) with estimates that were generated by splitting the CLASS scores into approximate quartiles (i.e., no assumption of linearity). Fourth, we explored models that included interactions between grade-level and the classroom quality measures to examine if the effect of classroom quality on child outcomes differed across grades.

Imputation for Covariates

Because our sample includes all students with at least one nonmissing CLASS score corresponding to one nonmissing behavioral or cognitive measure, some observations were missing lagged measures of classroom quality, behavioral problems, or achievement from the previous year. Across the five waves, approximately 8% of observations were missing lagged CLASS scores, 5% were missing lagged achievement scores, and 14% were missing lagged behavioral scores. Consequently, we used multiple imputations to account for missing data on the lagged control measures (as well as the age at outcome variable). For this process, we generated 25 multiply imputed data sets in Stata 16.0 (StataCorp, 2019) using the multivariate normal procedure. This imputation process included all analysis variables, including preschool measures of CLASS, achievement, and behavior. For the lagged control variables, we then averaged scores across the 25 imputed data sets to generate imputed measures of child age, lagged classroom quality (average CLASS score across the three domains), achievement (WJ-III composite), and behavior (SDQ Total Difficulties score), respectively.

Results

Descriptive Results

Table 1 presents descriptive statistics for the analysis sample, with sample characteristics shown for both the high- and lower risk groups. Children in the high-risk group were predominantly Black (70% Black; 28% White), the average income-to-needs ratio was 0.65, and their mothers completed an average of 11.86 years of schooling. In contrast, 36% of children in the lower risk group were Black (63% White), their average income-to-needs ratio was 2.19, and their mothers completed approximately 15.48 years of schooling. In each case, these group differences were statistically significant ($p < .001$).

CLASS domains and measures of achievement and behavior were averaged over the five elementary school waves for the descriptive results shown in Table 1. The gap between high- and lower risk students in achievement over elementary school was approximately .60 *SD*s ($p < .001$), and the behavior problems gap was approximately .35 *SD*s ($p < .001$). In Supporting Information (Tables S4 and S5), we present descriptive statistics for each wave included in the study, and these tables show that the gaps in achievement and behavior were relatively stable over the course of elementary school (though the behavioral gap grew somewhat during later grades).

Lower risk children were also more likely to be enrolled in higher quality classrooms as rated by the CLASS. Across the elementary school years, we observed that lower risk students were in classes that scored approximately 0.25 to 0.50 points higher (on the 1–7 scale; $p < .001$) on dimensions of the CLASS. However, we also observed substantial overlap in the distribution of CLASS scores between high- and lower risk students across the elementary school waves, suggesting that children in both groups experienced both higher and lower quality classrooms during elementary school (see Figures S2–S4). Table 1 also presents the proportion of students falling in the top quartile of each class domain, and we observed that the lower risk group was less likely to be in very high-quality classrooms. As Table 1 also reflects, these top quartile groups comprised less than 25% of the sample, as the scoring of the CLASS did not allow us to generate perfect quartile splits

Table 1
Descriptive Characteristics Disaggregated by Early Risk Factor

	High risk		Low risk		
	Top 25% on early risk factor		Bottom 75% on early risk factor		
	M	*SD* (within-child *SD*)	*M*	*SD* (within-child *SD*)	*p*-val of difference
CLASS domains (1–7)					
Emotional support	5.14	0.91 (0.77)	5.32	0.78 (0.66)	< .001
Classroom organization	5.06	0.86 (0.77)	5.27	0.77 (0.67)	< .001
Instructional support	2.87	1.05 (0.92)	3.17	1.00 (0.88)	< .001
Prop. in top quartile of CLASS domains					
Emotional support	0.20		0.22		.25
Classroom organization	0.13		0.18		< .001
Instructional support	0.16		0.23		< .001
Pre-K achievement and behavior					
WJ Achievement (*z*-score)	−0.43	0.91 (0.39)	0.14	0.99 (0.37)	< .001
SDQ Tot. Prob Beh (*z*-score)	0.27	1.04 (0.69)	-0.09	0.97 (0.58)	< .001
Demographic characteristics					
Female	0.50		0.50		.97
Black	0.70		0.36		< .001
Race—"Other"	0.02		0.01		.32
White	0.28		0.63		< .001
Income-to-needs ratio	0.65		2.19		< .001
Mother's years of completed schooling	11.86	2.21	15.48	2.18	< .001
Whether mother partnered	0.14		0.70		< .001
State—North Carolina	0.75		0.55		< .001
Unique observations	270		808		
Pooled observations	1,181		3,547		

Note. The "High Risk" group was composed of children in the top quartile of the early risk factor, and the "Low Risk" group was composed of children in the bottom 75% on the risk factor. For the CLASS, WJ, and SDQ descriptives presented, scores were averaged across all nonmissing observations between the kindergarten and Grade 5 waves of data collection. For each of the key analysis variables (CLASS scores, WJ achievement, and SDQ behavioral problems), we also present the average within-child *SD* in parentheses. The income-to-needs ratio, mother's education, and whether mother partnered measures were all averaged across the 6-, 15-, and 24-month waves of data collection. The "*p*-val of difference" column was generated from a series of regressions that regressed the listed variable on the indicator for whether the child was part of the "high-risk" group. CLASS = Classroom Assessment Scoring System; WJ = Woodcock–Johnson; SDQ, Strengths and Difficulties Questionnaire.

using Stata's "xtile" command (see Table S1 for scoring thresholds for each quartile group). Thus, these top quartile groups should be thought of as very high quality given the distributions observed in this sample. We also detail the results below from models that used alternative ways of calculating the classroom quartiles.

Fixed Effect Model Results

Table 2 presents results from child fixed effect models for achievement and behavioral problems. Each estimate was derived from a separate model. Estimates for the academic achievement composite score are shown in the top panel, whereas the estimates for the total problem behaviors score are shown in the bottom panel. For each set of results, we began with the basic fixed effects model (i.e., no lagged controls), before including the control for age, as well as lagged controls to examine whether classroom quality, achievement, or behavior from earlier periods affected our estimates.

For both achievement and behavioral problems, we observed no statistically significant effects of changes in *Emotional Support*. However, changes in *Classroom Organization* had small and statistically significant effects on both outcomes. A one-unit change in *Classroom Organization* was associated with a .03 *SD* ($p < .01$) change in academic achievement and a .07 *SD* ($p < .001$) reduction in behavioral problems. We found a similar effect on academic achievement for changes in *Instructional Support* ($b = .02$, $p < .01$). Reading across the columns in Table 2 also shows that including the lagged control variables

Table 2
Within-Child Associations Between CLASS Domains and Child Academic and Behavioral Outcomes Across Elementary School

	Academic achievement					
	Emotional support		Classroom organization		Instructional support	
	(1)	(2)	(3)	(4)	(5)	(6)
Continuous CLASS	0.01 (0.01)	0.00 (0.01)	0.03 (0.01)**	0.03 (0.01)**	0.02 (0.01)**	0.02 (0.01)*
Controls						
Child F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Grade level F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Achievement & behavior lag		Inc.		Inc.		Inc.
CLASS lag		Inc.		Inc.		Inc.
Age at outcome measurement		Inc.		Inc.		Inc.
Unique observations	1,070					
Total observations	4,613					
	Behavioral problems					
	Emotional support		Classroom organization		Instructional support	
	(7)	(8)	(9)	(10)	(11)	(12)
Continuous CLASS	−0.02 (0.02)	−0.02 (0.02)	−0.07 (0.02)***	−0.07 (0.02)***	−0.01 (0.01)	−0.02 (0.01)
Controls						
Child F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Grade level F.E.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Achievement & behavior lag		Inc.		Inc.		Inc.
CLASS lag		Inc.		Inc.		Inc.
Age at outcome measurement		Inc.		Inc.		Inc.
Unique observations	1,074					
Total observations	4,284					

Note. Robust standard errors were adjusted for child-level clustering and are presented in parentheses. All models included child fixed effects and grade level fixed effects. The second column in each set of estimates includes the 1-year lag of the achievement and behavioral measures, the 1-year lagged measure of the child's overall CLASS score (i.e., the average from the previous year of the three CLASS domains), and the child's age for each outcome measurement. The achievement and behavioral scores were both standardized within-grade. The "Unique Observations" row presents the number of children included in each model, and the "Total Observations" row presents the number of pooled observations (i.e., multiple observations per child) appearing in each model. The "Inc." marker designates when certain control variables were included in the model. CLASS = Classroom Assessment Scoring System.
*$p < .05$. **$p < .01$. ***$p < .001$.

had little effect on the results. In the Supporting Information (Table S6), we detail the results from models that tested for cross-lagged paths, and we found little indication that classroom quality in the previous period affected outcomes during the next year, nor did we find evidence that earlier child skills affected later classroom quality (nonsignificant point estimates ranged from .01 to .04 in magnitude).

Tables 3 and 4 show that the small average effects for the full sample mask sources of heterogeneity between the high- and lower risk groups. Across both the achievement (Table 3) and behavioral problems (Table 4) measures, we found that the benefits of moving into higher quality classes were generally larger and statistically significant for the high-risk group, whereas analogous changes had almost no effect on behavior or achievement for the lower risk group.

Beginning with the achievement models shown in Table 3, we found that chi-square tests of group differences indicated statistically significant differences between the high-risk and lower risk groups for each CLASS domain ($p < .01$). For the high-risk group, a one-unit change in *Classroom Organization* predicted a gain of approximately .06 *SD*s ($p < .01$) in achievement, and the effect for *Instructional Support* was similar ($b = .04$, $p < .01$). We observed no effect for lower risk children for *Emotional Support*,

Table 3
Within-Child Associations Between CLASS Domains and Academic Achievement Across Elementary School

	Academic achievement					
	Emotional support		Classroom organization		Instructional support	
	High risk	Lower risk	High risk	Lower risk	High risk	Lower risk
	(1)	(2)	(3)	(4)	(5)	(6)
Continuous CLASS	0.02 (0.02)	0.00 (0.01)	0.06 (0.02)**	0.02 (0.01)	0.04 (0.02)**	0.01 (0.01)
CLASS quartiles						
1st Quartile	ref.	ref.	ref.	ref.	ref.	ref.
2nd Quartile	0.13 (0.04)**	−0.01 (0.02)	0.07 (0.04)	0.00 (0.02)	0.10 (0.04)**	0.00 (0.02)
3rd Quartile	0.04 (0.05)	0.00 (0.02)	0.13 (0.04)**	0.01 (0.02)	0.13 (0.05)**	0.01 (0.02)
4th Quartile	0.07 (0.05)	0.01 (0.03)	0.11 (0.05)*	0.05 (0.03)	0.09 (0.05)*	0.02 (0.03)
p-val from test that all quartiles are equal	.02*	.92	.02*	.29	.01*	.78
p-val from test of model equivalence between groups	< .001***		< .001***		< .01**	
Full controls	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Unique observations	268	802	268	802	268	802
Total observations	1,150	3,463	1,150	3,463	1,150	3,463

Note. Robust standard errors were adjusted for student-level clustering and are presented in parentheses. All models included the full set of controls: child and year fixed effects, 1-year-lagged measures of achievement, behavior, and CLASS scores, and age at outcome measurement. The achievement score was standardized within-grade. In the top row, each estimate was generated from a separate child-fixed-effects regression that examined the effect of within-child changes in each respective CLASS domain on within-child changes in academic achievement across five measurement waves between kindergarten and fifth grade. Estimates in odd-numbered columns included only the high-risk group, and estimates in the even-numbered columns included only students in the lower risk group. The "Class Quartile" results were generated from separate models that tested distributional differences in the effect of CLASS scores. We generated a set of dummy variables indicating placement in quartiles of the multi-wave distribution of CLASS scores for each domain, and used the first quartile (i.e., the lowest quality classes) as the reference group in each model. The "*p*-val from test that all quartiles are equal" row presents *p*-values from post hoc tests that evaluated whether CLASS effects were equal across all quartiles of the CLASS distribution. The "*p*-val from test of model equivalence between groups" row presents *p*-values from chi-square tests evaluating whether each model significantly differed between the two risk groups. This test was executed for the "class quartile" models using the "suest" command in Stata. CLASS = Classroom Assessment Scoring System.
*$p < .05$. **$p < .01$. ***$p < .001$.

and across all three CLASS domains, we found no statistically significant effects for the lower risk group.

In the second panel of Table 3, we tested whether the benefits of changes in classroom quality were driven by threshold effects. In these models, we used quartile indicators for each respective domain of the CLASS, with classrooms falling in the bottom 25% of the distribution serving as the comparison group. Focusing on the high-risk group, we found evidence of nonlinearity. For *Classroom Organization* and *Instructional Support*, the effect on achievement was primarily driven by moving out of a classroom in the bottom quartile (*Classroom Organization:* $p < .05$; *Instructional Support*: $p < .05$). Moving to higher quartiles of classroom quality did not provide additional statistically significant benefits for either domain.

Table 4 presents results for the SDQ measure of behavioral problems. Here, we observed statistically significant effects for the *Classroom Organization* domain for both groups. A one-unit change in *Classroom Organization* predicted a .13 *SD* ($p < .001$) reduction in behavioral problems for high-risk children, and the effect for the lower risk children was statistically significant, but smaller ($b = -.05$, $p < .05$). We also observed that a 1-unit change in *Emotional Support* predicted a .08 ($p < .05$) *SD* reduction in behavioral problems for the highest risk students. We observed no statistically significant effects for *Instructional Support*.

For the behavioral-problems models, we again observed evidence of threshold effects for the highest risk students, as the *Classroom Organization* effect was driven by the highest quality classrooms. Moving out of a bottom quartile class and into a top quartile class on the *Classroom Organization* domain predicted a large reduction in behavioral problems ($b = -.38$, $p < .001$) for high-risk students, and post hoc tests revealed that moving into a top quartile class provided statistically significant benefit over moving into a third quartile class ($p < .01$), though the benefit over the second quartile was only marginally significant ($p = .06$).

Table 4
Within-Child Associations Between CLASS Domains and Problem Behaviors Across Elementary School

	Behavioral problems					
	Emotional support		Classroom organization		Instructional support	
	High risk	Lower risk	High risk	Lower risk	High risk	Lower risk
	(1)	(2)	(3)	(4)	(5)	(6)
Continuous CLASS	−0.08 (0.04)*	0.00 (0.02)	−0.13 (0.04)***	−0.05 (0.02)*	−0.04 (0.03)	−0.01 (0.02)
CLASS quartiles						
1st Quartile	ref.	ref.	ref.	ref.	ref.	ref.
2nd Quartile	0.00 (0.07)	−0.08 (0.04)	−0.21 (0.07)**	−0.09 (0.04)*	−0.12 (0.07)	−0.08 (0.04)
3rd Quartile	0.04 (0.08)	−0.05 (0.04)	−0.11 (0.08)	−0.14 (0.04)***	−0.08 (0.08)	−0.03 (0.04)
4th Quartile	−0.15 (0.09)	0.00 (0.05)	−0.38 (0.10)***	−0.10 (0.05)*	−0.04 (0.09)	−0.08 (0.04)
p-val from test that all quartiles are equal	.12	.13	< .01**	.01*	.36	.15
p-val from test of model equivalence between groups	< .01**		< .01**		.02*	
Full controls	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Unique observations	269	805	269	805	269	805
Total observations	1,071	3,213	1,071	3,213	1,071	3,213

Note. See Table 3 note. For these models, all estimates were generated using the within-grade standardized "Total Difficulties" scale of the SDQ as the dependent variable. All models included full controls: child and year fixed effects, 1-year-lagged measures of achievement, behavior, and CLASS scores, and age at outcome measurement. CLASS = Classroom Assessment Scoring System; SDQ = Strengths and Difficulties Questionnaire.
*$p < .05$. **$p < .01$. ***$p < .001$.

In Table 5, we present results from models for high-risk children that included the three CLASS domains simultaneously, allowing us to test the unique effect of within-child changes in each CLASS domain. These results again showed that much of the classroom quality benefit was carried by changes in *Classroom Organization* for the high-risk students, as we found statistically significant effects of *Classroom Organization* on both achievement ($b = .07$, $p < .01$) and behavior ($b = -.16$, $p < .001$). When considering the CLASS domains simultaneously, we observed no reliable effects for *Instructional Support*, though we observed a negative effect for *Emotional Support* on academic achievement ($b = -.05$, $p < .05$). However, this finding should be interpreted cautiously, as it was only observed when all three CLASS domains were included in the same model, and we observed no statistically significant relations between the continuous measure of *Emotional Support* and achievement for both the high- and lower risk samples when CLASS domains were considered independently.

Because we found that effects for both achievement and behavior were driven largely by changes in the *Classroom Organization* domain, we present results in Table 5 from models that tested interactions between *Classroom Organization* and the early risk factor (results for the other domains are presented in Supporting Information). For achievement, we observed a statistically significant continuous interaction ($b = .03$, $p < .01$), with both the classroom quality and risk measures included as continuous variables. This interaction suggests that moving into classrooms with higher levels of *Classroom Organization* produced increasing returns to achievement as children experienced more early environmental risk. In contrast, we found no continuous interaction for reductions in behavioral problems, but found a substantial interaction between the high-risk indicator and the top quartile of the *Classroom Organization* distribution ($b = -.22$, $p < .05$). This result further suggested that large improvements in *Classroom Organization* led to meaningful reductions in behavioral problems for the highest risk students.

Finally, we explored whether classroom quality effects were consistent across elementary school years (results presented in Tables S7 and S8). For academic achievement, we found little indication that effects differed between grade-levels, as all tested interactions were statistically nonsignificant. Thus, the benefits of classroom quality for achievement did not differ between the early and later grades, and this pattern was consistent for both high- and lower risk children. For behavioral problems, we again observed fairly consistent effects

Table 5
Interactions Between Early Risk and Classroom Organization

	Achievement			Behavioral problems		
	High risk	Full sample		High risk	Full sample	
	(1)	(2)	(3)	(4)	(5)	(6)
Emotional support (cont.)	−0.05 (0.03)*	−0.03 (0.01)*	−0.03 (0.01)*	0.03 (0.05)	0.04 (0.02)	0.03 (0.02)
Classroom organization (cont.)	0.07 (0.03)**	0.03 (0.01)*	—	−0.16 (0.05)***	−0.11 (0.02)***	—
Instructional support (cont.)	0.03 (0.02)	0.02 (0.01)	0.01 (0.01)	0.02 (0.03)	0.01 (0.02)	0.00 (0.02)
CLASS organization categories						
2nd Quartile			0.01 (0.03)			−0.12 (0.04)**
3rd Quartile			0.02 (0.03)			−0.18 (0.05)***
4th Quartile			0.06 (0.03)			−0.15 (0.05)**
Interactions with risk						
Class Org. (cont.) × Risk (cont.)		0.03 (0.01)**			0.00 (0.02)	
Class Org. 2nd Quartile × High Risk			0.06 (0.05)			−0.07 (0.08)
Class Org. 3rd Quartile × High Risk			0.11 (0.05)*			0.05 (0.08)
Class Org. 4th Quartile × High Risk			0.04 (0.05)			−0.22 (0.10)*
Full controls	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Unique observations	268	1,070	1,070	269	1,074	1,074
Total observations	1,150	4,613	4,613	1,071	4,284	4,284

Note. Robust standard errors were adjusted for child-level clustering and are presented in parentheses. All models include the full set of controls: child and year fixed effects, 1-year-lagged measures of achievement, behavior, and CLASS scores, and age at outcome measurement. Columns 1 through 3 included the standardized measure of achievement as the dependent variable, and Columns 4 through 6 included the standardized measure of problem behaviors as the dependent variable. Columns 1 and 4 include the high-risk sample only. In Columns 2 and 5, we tested the interaction between the continuous measure of *Classroom Organization* and the continuous early risk factor. In Columns 3 and 6, we tested interactions between the quartiles of the *Classroom Organization* distribution and the indicator for whether a student was in the high-risk group. CLASS = Classroom Assessment Scoring System.
*$p < .05$. **$p < .01$. ***$p < .001$.

across grade levels for the highest risk students. However, for the lower risk students, we found that improvements in classroom quality slightly reduced behavioral problems in early grades, but led to increases in behavioral problems in fifth grade. While potentially interesting, conclusions based on these results should be tempered given that they were not hypothesized, and average quality effects across waves were generally close to zero for the lower risk children.

Sensitivity Tests

The supplementary material presents the estimates from a host of additional analyses that extend our key results. We briefly describe these supporting analyses here, and provide full details in Supporting Information.

In Tables S9 and S10, we present key results using standardized measures of classroom quality, which allows coefficients to be interpreted as effect sizes. These standardized results are, in most cases, nearly identical to the results reported in the main text because the raw-unit *SD*s for the various CLASS domains were all close to "1." We also present interactions between the risk factor and the measures of *Emotional Support* and *Instructional Support* (Table S11; compare with results shown in Table 5). We found few statistically significant interactions for either domain.

Next, we tested whether our results were robust to alternative measures of early environmental risk (Tables S12 and S13). Specifically, we assessed whether CLASS effects were larger for students who had higher levels of early exposure to domestic violence, or for those who came from homes scoring in the bottom quartile of an observational assessment of the early home environment. Across both of these alternative early environmental risk measures, we found that effects tended to be larger for the higher risk students.

Recall that our quantile measures of classroom quality were derived by assessing quality across the entire distribution of CLASS scores from all of the elementary school waves. In Table S14, we present results from models that calculated CLASS quantiles separately for each wave of the study, which led to different score thresholds used to determine the quartile groups in each wave. As Table S14 reflects, results were generally similar to those

shown in our key tables (though the effect for the top quartile for *Classroom Organization* group in the behavioral problems model was smaller by about .10 *SD*s).

Finally, Tables S15 and S16 present results for the subscores used to create our composite measures of achievement and behavior. For achievement, we found that results were largest for the two measures of reading—*Letter Word Identification* and *Passage Comprehension*. We found smaller effects for *Applied Problems*, and surprisingly, we observed negative point estimates for the *Picture Vocabulary* measure. For the SDQ measures, we found that results were largest for the *Peer Problems* and *Conduct Problems* scores, and results were smaller, but in the same direction for *Hyperactivity* and *Emotional Symptoms*. However, it should be noted that these results were estimated with slightly less precision, and many of the subscore coefficients were not statistically significantly different from one another.

Discussion

Links between measures of classroom quality and children's developmental outcomes have been reported across multiple samples (e.g., Mashburn et al., 2008; Rimm-Kaufman et al., 2015; Vernon-Feagans et al., 2019), yet causal identification in this literature has been limited due to the difficulty of controlling for unobserved factors that lead children to select into classroom environments. This study used child fixed effects models to examine associations between within-child changes in classroom quality and within-child changes in achievement and behavioral problems across elementary school grades. As expected, our estimates were in the smaller range of effect sizes reported in previous correlational work (see Perlman et al., 2016 meta-analysis). Across models, we found that changes in classroom quality tended to have small and statistically nonsignificant effects on achievement and behavior. However, we found that for the highest risk students, moving into classrooms that provided more organization and better behavioral management had positive, but modest, effects on children's academic achievement and behavioral adjustment. For these same children, moving into classrooms with higher quality instruction boosted achievement, and moving into classrooms that provided more emotional support also reduced behavioral problems. When all three domains of classroom quality were considered simultaneously, we found that classroom quality effects on both achievement and behavior were uniquely driven by changes in *Classroom Organization*.

Interestingly, our analyses of quality thresholds for high-risk students suggested that the academic benefits of quality changes were driven by moving out of the lowest quality classrooms and in to mid-quality classrooms. Perhaps surprisingly, we found no additional benefit on achievement for moving into classes that fell in the upper part of the distribution on *Classroom Organization* and *Instructional Support*. This suggests that high-risk students experienced achievement growth as a result of leaving classrooms that were relatively very poorly managed. However, it should be noted that most classrooms in this study were not rated particularly favorably on the *Instructional Support* domain, as the average classroom was rated at 3.10 on the "1" to "7" scale, and classrooms in the top quartile of the distribution had average *Instructional Support* scores of only 4.53. Indeed, studies of other samples have reported similarly low averages for the *Instructional Support* domain (e.g., Burchinal et al., 2010). Yet, it remains possible that students could have benefited from moving into very high-quality classrooms that were simply outside the range of classroom experiences observed in this study.

We found a different distributional pattern for the association between classroom quality and behavioral problems. For the highest risk students, the effect of moving into a well-managed classroom on child behavior was strongest at the very top of the *Classroom Organization* distribution (i.e., classroom scorings above a "6"), and this was primarily driven by the reduction in problems with social adjustment and externalizing behavior (see Supporting Information). In other words, moving out of a very poorly managed class into an especially well-managed class led to a sizable reduction in behavioral problems for the highest risk students (.38 *SDs*), whereas moving to classrooms in the middle of the distribution led to smaller reductions in behavioral problems. Indeed, the effect of moving from the bottom to top quartile in *Classroom Organization* accounted for approximately all of the gap in behavioral problems between high- and lower risk students, and these results held even when controlling for the other domains of classroom quality. Thus, efforts to improve the classroom management of teachers working in highly disadvantaged contexts may provide an effective means of behavioral intervention for children from severely impoverished homes (Raver et al., 2009). However, our findings also suggest that in order to produce meaningful change in behavior, efforts to

improve teacher organization strategies would need to produce very well-managed and organized classrooms. Middle-of-the-road improvements may not be enough.

For the highest risk students in the sample, our analyses also suggested that improvements in classroom quality largely produced consistent effects across the developmental periods covered in elementary school. This may be surprising, as many have argued that early educational experiences may be most crucial (e.g., Ramey & Ramey, 1998). However, these findings suggest that each year of a child's elementary school experience is important for providing necessary support for both behavioral and academic growth, and these results further bolster calls by many early childhood researchers to place more attention on the quality of children's educational experiences throughout elementary school (Stipek, Franke, Clements, Farran, & Coburn, 2017).

Perhaps surprisingly, our analyses showed that for lower risk students in the sample, domains of classroom-wide quality captured by the CLASS did not relate to broad measures of achievement and behavioral functioning. This suggests that our efforts to measure essential elements of children's classroom experiences could still be improved. Some have argued that observational instruments like the CLASS may miss important elements of children's classroom experiences, especially those pertaining to the learning opportunities of minority children in the classroom (Curenton et al., 2020). We found that CLASS effects were largest for the high-risk group, which was predominantly comprised of Black children (70%). Yet, future work should continue to examine which specific classroom processes lead to the largest benefits for the most disadvantaged learners. Indeed, the CLASS attempts to capture general classroom quality for all students in a given classroom, and perhaps researchers could capture more meaningful variation if they relied on individualized measures of students' experiences (see Kim et al., 2019). However, it should be noted that a previous study of teacher-rated relationship quality with individual students using within child-modeling also found limited and modest effects on direct assessments of achievement (Maldonado-Carreño & Votruba-Drzal, 2011).

However, in the important case of the highest risk students, improvements in *Classroom Organization* and *Instructional Support* did appear to promote these students' school achievement and positive behaviors. The fact that improvements in these domains of classroom quality produced benefits on achievement and behavior for the most severely disadvantaged students (i.e., average income-to-needs ratio = 0.65) provides further evidence that high-quality school environments might serve as a protective factor against negative developmental influences that stem from growing up in highly impoverished homes during early life (Dearing et al., 2009). Our findings echo other recent work in the FLP sample by Vernon-Feagans et al. (2019) that reported that the accumulation of high-quality classroom experiences over 4 years had the strongest benefit for children who began elementary school with low literacy scores.

Moreover, these findings were further supported by results shown in the Supporting Information, as we found larger classroom quality effects for children whose mothers reported high levels of exposure to domestic violence, and we observed a similar pattern for children whose early home environments were rated as low-quality on a home observational assessment. Interestingly, the lack of findings for the lower risk sample could suggest that environmental experiences act as substitutes for one another. In other words, children from more supportive home environments may have less to gain from higher quality classroom experiences, whereas children from high-risk environments stand to benefit greatly from time spent in higher quality classrooms (see work on "functionally equivalent environments," described in McCartney & Berry, 2009; and recent work on "environmental substitutability" by Bailey, Duncan, Cunha, Foorman, & Yeager, 2020).

Limitations

Several important limitations should be noted. First, our estimates were not generated using nationally representative data, so generalizations to other samples should be made with caution. Second, our results for behavioral problems were generated using teacher-reported measures of students' behavior. Thus, we cannot rule out whether teachers with better-managed classrooms simply perceived their students as being better behaved, even if these students showed no actual behavioral change. Third, our analysis only covered the early elementary grades up to Grade 5. It is an open question whether classroom quality as measured by the CLASS would continue to relate to student achievement and behavioral problems during adolescence. Fourth, we cannot fully rule out whether bidirectional relations exist between children's

achievement or behavior and classroom quality. Certainly, teachers respond to the skills and capacities of their children. Yet, our results were consistent when controlling for achievement and behavior during the previous year, and these controls should adjust for any changes in teaching caused by children's initial skills.

Finally, as with most correlational analyses, our results were likely affected by omitted variables bias and measurement error in our independent variables. Observers only rated each classroom over two 30-min cycles, likely leading to measurement error that could bias effects toward zero. Conversely, unmeasured changes in child and family characteristics could still drive selection into higher quality classes (see critique in Rothstein, 2009), and such effects would likely bias our estimates upwards. Although our controls for lagged measures should limit bias due to time-varying factors, we cannot fully rule out such possibilities.

Conclusion

Our results suggest that the classroom quality dimensions captured by the CLASS likely have some limited effects on student achievement and behavioral problems during elementary school for children exposed to high levels of poverty during early childhood. Indeed, children from the most disadvantaged homes would likely benefit most from efforts to improve classroom quality—particularly programs that make substantial improvements to classroom organization and behavioral management.

References

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment*. Hillsdale, NJ: Erlbaum. Retrieved from http://parentalalienationresearch.com/PDF/2015ainsworth.pdf

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system—Secondary. *School Psychology Review*, *42*, 76–98. https://doi.org/10.1080/02796015.2013.12087492

Ansari, A., & Pianta, R. C. (2018). Variation in the long-term benefits of child care: The role of classroom quality in elementary school. *Developmental Psychology*, *54*, 1854–1867. https://doi.org/10.1037/dev0000513

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

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fadeout of educational intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest*, *21*, 55–97. https://doi.org/10.1177/1529100620915848

Bailey, D. H., Duncan, G. J., Watts, T. W., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist*, *73*, 81–94. https://doi.org/10.1037/amp0000146

Berry, D., Blair, C., Willoughby, M., Garrett-Peters, P., Vernon-Feagans, L., Mills-Koonce, W. R.; Family Life Project Key Investigators. (2016). Household chaos and children's cognitive and socio-emotional development in early childhood: Does childcare play a buffering role? *Early Childhood Research Quarterly*, *34*, 115–127. https://doi.org/10.1016/j.ecresq.2015.09.003

Berry, D., & Willoughby, M. T. (2017). On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse. *Child Development*, *88*, 1186–1206. https://doi.org/10.1111/cdev.12660

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology*, *66*, 711–731. https://doi.org/10.1146/annurev-psych-010814-015221

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). Narrowing the gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management*, *27*, 793–818. https://doi.org/10.1002/pam.20377

Broekhuizen, M. L., Mokrova, I. L., Burchinal, M. R., Garrett-Peters, P. T.; Family Life Project Key Investigators. (2016). Classroom quality at pre-kindergarten and kindergarten and children's social skills and behavior problems. *Early Childhood Research Quarterly*, *36*, 212–222. https://doi.org/10.1016/j.ecresq.2016.01.005

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*, 166–176. https://doi.org/10.1016/j.ecresq.2009.10.004

Burchinal, M., & Willoughby, M. (2013). IV. Poverty and associated social risks: Toward a cumulative risk framework. *Monographs of the Society for Research in Child Development*, *78*, 53–65. https://doi.org/10.1111/mono.12050

Burchinal, M., Xue, Y., Auger, A., Tien, H. C., Mashburn, A., Peisner-Feinberg, E., … Tarullo L. (2016). Testing for quality thresholds and features in early care and education. *Monographs of the Society for Research in Child Development*, *81*, 46–63. https://doi.org/10.1111/mono.12238

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*, 292–309. https://doi.org/10.1080/10409289.2012.665760

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

Curenton, S. M., Iruka, I. U., Humphries, M., Jensen, B., Durden, T., Rochester, S. E., … Kinzie, M. B. (2020). Validity for the Assessing Classroom Sociocultural Equity Scale (ACSES) in early childhood classrooms. *Early Education and Development*, *31*, 269–288. https://doi.org/10.1080/10409289.2019.1611331

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

Dearing, E., McCartney, K., & Taylor, B. A. (2009). Does higher quality early child care promote low-income children's math and reading achievement in middle childhood? *Child Development*, *80*, 1329–1349. https://doi.org/10.1111/j.1467-8624.2009.01336.x

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*, 1337–1345. https://doi.org/10.1097/00004583-200111000-00015

Goodman, R., & Scott, S. (1999). Comparing the strengths and difficulties questionnaire and the child behavior checklist: is small beautiful? *Journal of Abnormal Child Psychology*, *27*, 17–24. https://doi.org/10.1023/A:1022658222914

Guerrero-Rosada, P., Weiland, C., McCormick, M., Hsueh, J., Sachs, J., Snow, C., & Maier, M. (2021). Null relations between CLASS scores and gains in children's language, math, and executive function skills: A replication and extension study. *Early Childhood Research Quarterly*, *1*, 1–121. https://doi.org/10.1016/j.ecresq.2020.07.009

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

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, *76*, 949–967. https://doi.org/10.1111/j.1467-8624.2005.00889.x

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

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*, 1900–1902. https://doi.org/10.1126/science.1128898

Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, *63*, 467–490. https://doi.org/10.1111/ajps.12417

Jennings, P. A., & Greenberg, M. T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research*, *79*, 491–525. https://doi.org/10.3102/0034654308325693

Kersten, P., Czuba, K., McPherson, K., Dudley, M., Elder, H., Tauroa, R., & Vandal, A. (2016). A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire. *International Journal of Behavioral Development*, *40*, 64–75. https://doi.org/10.1177/0165025415570647

Kim, H., Cameron, C. E., Kelly, C. A., West, H., Mashburn, A. J., & Grissmer, D. W. (2019). Using an individualized observational measure to understand children's interactions in underserved kindergarten classrooms. *Journal of Psychoeducational Assessment*, *37*, 935–956. https://doi.org/10.1177/0734282918819579

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*, 547–588. https://doi.org/10.3102/0034654318759268

Li, W., Farkas, G., Duncan, G. J., Burchinal, M. R., & Vandell, D. L. (2013). Timing of high-quality child care and cognitive, language, and preacademic development. *Developmental Psychology*, *49*, 1440–1451. https://doi.org/10.1037/a0030613

Maldonado-Carreño, C., & Votruba-Drzal, E. (2011). Teacher–child relationships and the development of academic and behavioral skills during elementary school: A within-and between-child analysis. *Child Development*, *82*, 601–616. https://doi.org/10.1111/j.1467-8624.2010.01533.x

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

McCartney, K., & Berry, D. (2009). Whether the environment matters more for children in poverty. In K. McCartney & R. A. Weinberg (Eds.), *Experience and development: A festschrift in honor of Sandra Wood Scarr* (pp. 99–124). Hove, UK: Psychology Press.

McCormick, M. P., Cappella, E., O'Connor, E. E., & McClowry, S. G. (2015). Social-emotional learning and academic achievement: Using causal methods to explore classroom-level mechanisms. *AERA Open*, *1*(3). https://doi.org/10.1177/2332858415603959

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., & Yazejian, N. (2003). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development*, *72*, 1534–1553. https://doi.org/10.1111/1467-8624.00364

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

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*, 365–397. https://doi.org/10.3102/0002831207308230

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*, 109–119. https://doi.org/10.3102/0013189X09332374

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*, 956–975. https://doi.org/10.1080/10409289.2017.1319783

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Manual K-3*. Baltimore, MD: Paul H. Brookes Publishing. Retrieved from https://psycnet.apa.org/record/2007-18799-000

Ramey, C. T., & Ramey, S. L. (1998). Early intervention and early experience. *American Psychologist*, *53*, 109–120. https://doi.org/10.1037/0003-066X.53.2.109

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Metzger, M. W., & Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *77*, 302–316. https://doi.org/10.1037/a0015302

Rea, D., & Burton, T. (2020). New evidence on the Heckman curve. *Journal of Economic Surveys*, *34*, 241–262. https://doi.org/10.1111/joes.12353

Rimm-Kaufman, S. E., Baroody, A. E., Larsen, R. A., Curby, T. W., & Abry, T. (2015). To what extent do teacher–student interaction quality and student gender contribute to fifth graders' engagement in mathematics learning? *Journal of Educational Psychology*, *107*, 170–185. https://doi.org/10.1037/a0037252

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*, 958–972. https://doi.org/10.1037/a0015861

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*, 537–571. https://doi.org/10.1162/edfp.2009.4.4.537

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*, 68. https://doi.org/10.1037/0003-066X.55.1.68

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype→environment effects. *Child Development*, *54*, 424–435. https://doi.org/10.2307/1129703

StataCorp. (2019). *Stata statistical software: Release 16*. College Station, TX: StataCorp LLC.

Stipek, D., Franke, M., Clements, D., Farran, D., & Coburn, C. (2017). PK-3: What does it mean for instruction? Social policy report. *Society for Research in Child Development*, *30*. https://doi.org/10.1002/j.2379-3988.2017.tb00087.x

Vernon-Feagans, L., Cox, M.; FLP Key Investigators. (2013). The family life project: An epidemiological and developmental study of young children living in poor rural communities. *Monographs of the Society for Research in Child Development*, *78*, 1–150. Retrieved from https://www.jstor.org/stable/43773265

Vernon-Feagans, L., Mokrova, I. L., Carr, R. C., Garrett-Peters, P. T., & Burchinal, M. R.; Family Life Project Key Investigators. (2019). Cumulative years of classroom quality from kindergarten to third grade: Prediction to children's third grade literacy skills. *Early Childhood Research Quarterly*, *47*, 531–540. https://doi.org/10.1016/j.ecresq.2018.06.005

Weiland, C., Ulvestad, K., Sachs, J., & Yoshikawa, H. (2013). Associations between classroom quality and children's vocabulary and executive function skills in an urban public prekindergarten program. *Early Childhood Research Quarterly*, *28*, 199–209. https://doi.org/10.1016/j.ecresq.2012.12.002

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

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M. C., Weiland, C., … Arbour, M. C. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Developmental Psychology*, *51*, 309–322. https://doi.org/10.1037/a0038785

Zachrisson, H. D., & Dearing, E. (2015). Family income dynamics, early childhood education and care, and early child behavior problems in Norway. *Child Development*, *86*, 425–440. https://doi.org/10.1111/cdev.12306

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Appendix S1. The supplementary file contains further details regarding study methods, as well as additional results from supplementary analyses