Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1395

Classroom Composition Affects Teacher Performance Ratings

William Delgado
University of Chicago

Lauren Sartain
University of North Carolina at Chapel Hill

Teacher evaluations should reflect teaching performance rather than the characteristics of the students assigned to a teacher. Exploiting naturally occurring year-to-year variation in classroom composition within teachers, this paper examines whether teacher performance ratings assigned by evaluators and students are influenced by classroom context. We find that teachers with higher-achieving and less disruptive students, holding constant the teacher and school, receive systematically higher performance ratings. These effects are robust across model specifications, placebo tests, and multiple dimensions of teaching practice. By contrast, classroom demographics show no consistent association with performance ratings. A policy that adjusts evaluator scores for classroom characteristics, analogous to value-added models, increases the relative ranking of Black teachers by 8 percentage points, highlighting equity impacts of considering classroom context.

VERSION: April 2026

Suggested citation: Delgado, William, and Lauren Sartain. (2026). Classroom Composition Affects Teacher Performance Ratings. (EdWorkingPaper: 26-1395). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/2c7t-mt76

Classroom Composition Affects Teacher Performance Ratings *

William Delgado † Lauren Sartain ‡

April 2026

Abstract

Teacher evaluations should reflect teaching performance rather than the characteristics of the students assigned to a teacher. Exploiting naturally occurring year-to-year variation in classroom composition within teachers, this paper examines whether teacher performance ratings assigned by evaluators and students are influenced by classroom context. We find that teachers with higher-achieving and less disruptive students, holding constant the teacher and school, receive systematically higher performance ratings. These effects are robust across model specifications, placebo tests, and multiple dimensions of teaching practice. By contrast, classroom demographics show no consistent association with performance ratings. A policy that adjusts evaluator scores for classroom characteristics, analogous to value-added models, increases the relative ranking of Black teachers by 8 percentage points, highlighting equity impacts of considering classroom context.

Keywords: teacher quality, subjective performance ratings, classroom observations, student surveys, classroom characteristics, teacher characteristics

JEL codes: I21, J45, M50

*The authors are extremely grateful to Andrew Zou for outstanding research assistance in early versions of the paper. We also thank Elaine Allensworth, John Q. Easton, the staff at the UChicago Consortium on School Research, and the staff at Chicago Public Schools, particularly the Talent Office, for helping us use the administrative data and better understand the policy context. We are extremely grateful to Seth Zimmerman, Dan Black, Damon Jones, Stephen Raudenbush, and Susan E. Mayer for helpful comments and numerous conversations. We are also grateful to Josh Goodman, Marcus Winters, Andrew Bacher-Hicks, Olivia Chi, Kirsten Slungaard Mumma, Alexis Orellana and colleagues and students at Boston University Wheelock College of Education and Human Development for comments and conversations. We thank seminar participants at AEFP, SREE, Economics of Racism Seminar, and Northeastern Economics of Education Seminar. We are also grateful for comments from anonymous referees. The authors gratefully acknowledge funding for this research from the Spencer Foundation. Any errors are our own.

†UChicago Consortium on School Research, Crown Family School of Social Work, Policy, and Practice, University of Chicago, Chicago, IL 60637 (email: wdelgado@uchicago.edu).

‡School of Education, University of North Carolina at Chapel Hill, Peabody Hall, Chapel Hill, NC 27514 (email: lsartain@unc.edu).

1 Introduction

Amid concerns that traditional teacher evaluation systems failed to identify low-performing teachers for remediation or removal, the federal Race to the Top (RTTT) initiative spurred districts across the United States to adopt more rigorous teacher evaluation policies. Most large districts now use evaluation systems that combine classroom observation rubrics with measures of student growth (Steinberg and Donaldson, 2016). Because these systems carry high stakes for teachers' careers, understanding whether subjective performance ratings reflect instructional quality or classroom context beyond teachers' control is policy relevant.

While value-added measures (VAMs) have received considerable attention and have been shown to provide unbiased estimates of teachers' causal contributions to student learning (Kane and Staiger, 2008; Chetty et al., 2014a), they are only available for a minority of teachers and remain controversial due to concerns about reliability and bias (Rothstein, 2009, 2017; Papay, 2011). In practice, classroom observations—conducted by school administrators using standardized rubrics—receive the most weight in teacher evaluations, supplemented in some districts by student surveys. These subjective measures predict student achievement (Harris and Sass, 2014; Jacob and Lefgren, 2008; Sartain et al., 2011), but they may also be sensitive to factors unrelated to teacher effectiveness.

Teachers themselves have raised concerns that classroom observation ratings are shaped by the characteristics of the students they teach. For instance, teachers report it is harder to earn top ratings when serving larger proportions of students with behavioral challenges, special education needs, or limited English proficiency. Evaluators' expertise and the cultural fit of the rubric may further influence scores. A persistent finding in the literature is that teachers serving students with lower prior achievement or from disadvantaged backgrounds tend to receive lower observation scores (Chaplin et al., 2014; Whitehurst et al., 2014; Sporte and Jiang, 2016). This correlation presents a key identification problem: it could be a result of non-random sorting of teachers to schools and classrooms, or it could reflect a causal effect of classroom context on ratings. Causal channels could include teaching being genuinely more difficult in more challenging settings, or evaluator bias, where raters subconsciously penalize or reward teachers based on the students they serve. At stake is whether these measures reflect teaching practice or classroom context.

This paper examines whether classroom composition influences teacher performance ratings. We ask: (i) To what extent are administrator observation ratings and student survey reports affected by the characteristics of students in the classroom? and (ii) How would adjusting observation ratings for student characteristics change the distribution of teacher rankings, and which teachers would be most affected? The classroom factors we study include student academic and behavioral measures (prior-year test scores, GPA, attendance, suspensions, and grade repetition), as well as student demographics (gender, race/ethnicity, free or reduced-price lunch eligibility, and special education status), and class size.

Using five years of administrative data from Chicago Public Schools (CPS)—the third-largest school district in the U.S. at the time of the data collection—we employ a teacher fixed-effect design to address identification challenges. This quasi-experimental design leverages natural year-to-year variation in classroom composition for the same teacher, allowing us to isolate the causal effect of student characteristics while controlling for time-invariant teacher quality and sorting.

Our findings provide evidence of a causal effect of classroom characteristics on teacher performance ratings. We find that teachers receive significantly higher ratings in years when they teach higher-achieving students with fewer behavioral challenges, even when they remain in the same school. Specifically, a 1 SD increase in a constructed classroom quality index (based on baseline academic and behavioral measures) leads to a 0.07 SD increase in classroom observation ratings and a 0.13 SD increase in student survey scores. These effects appear across multiple dimensions of practice, including classroom management and instruction. By contrast, we find that classroom demographic composition does not consistently predict ratings. Results are robust across alternative specifications, placebo tests, and subsample analyses. For example, we test whether current classroom composition predicts past or future teacher ratings once teacher fixed effects are included, and find null effects, supporting exogeneity of within-teacher composition shocks. Furthermore, our exploration of mechanisms indicates that these classroom effects are not reflected in outcome-based measures of teacher productivity (i.e., student growth measures), pointing to evaluator bias or other factors rather than genuine changes in teacher effectiveness as the likely driver.

Finally, a policy simulation shows that adjusting ratings for classroom characteristics would meaningfully change teacher rankings, with Black teachers benefiting the most. Their average ranking would improve by about 8 percentile points relative to non-Black peers.

This study contributes to the literature on teacher effectiveness and evaluation by providing large-scale, quasi-experimental evidence that classroom context systematically affects subjective performance ratings. Our findings highlight both measurement validity concerns and equity implications for teacher evaluation policies and suggest possible ways to address them.

2 Related Literature

The importance of high-quality teachers for student achievement and later-life outcomes is well-established (Sanders and Horn, 1998; Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Chetty et al., 2014b). Recognizing this, policymakers have increasingly focused on developing more effective teacher evaluation systems. Spurred by federal initiatives like Race to the Top (RTTT), states and districts moved to implement evaluation systems using multiple measures of teacher practice, which heavily emphasize rubric-based classroom observations conducted by school administrators (Doherty and Jacobs, 2013; Steinberg and Donaldson, 2016). While much academic and public debate has centered on the validity of student growth measures like VAMs, classroom observations remain the cornerstone of most evaluation systems, carrying the most weight in final ratings and applying to all teachers, unlike VAMs which are only available for tested grades and subjects (Kane and Staiger, 2008; Rothstein, 2009; Papay, 2011; Chetty et al., 2014a).

Evidence shows that both subjective and objective measures of teacher effectiveness predict future student achievement (Jacob and Lefgren, 2008; Rockoff and Speroni, 2011), and strong teaching can be especially important for lower-achieving students (Aaronson et al., 2007). Much of the validation research of subjective performance ratings has compared observation ratings to VAMs. A study using findings from Chicago's Excellence in Teaching Pilot found that, on average, teachers with higher classroom observation ratings had significantly higher value-added measures (Sartain et al., 2011). Students also showed the most growth in classrooms of highly rated teachers, and the least growth in classrooms of poorly rated teachers. These findings are supported by the five observation instruments used in the Measures of Effective Teaching (MET) project, which were positively associated with student achievement gains (Kane and Staiger, 2012).

A central challenge to the validity of these observation ratings is their consistent correlation with classroom composition. Several studies find that teachers serving students with lower prior

achievement, limited English proficiency, or from low-income backgrounds tend to receive lower observation scores (Chaplin et al., 2014; Whitehurst et al., 2014; Sporte and Jiang, 2016). For example, Whitehurst et al. (2014) found that teachers of low-achieving students were nearly four times less likely to be rated in the top quintile compared to teachers of high-achieving students. These correlations raise concerns that, rather than reflecting instructional quality, ratings may be conflated with the classroom context in which teaching occurs—a context often outside a teacher's control.

Interpreting this correlation presents a key identification problem. One explanation is the non-random sorting of teachers to schools and classrooms. More experienced and credentialed teachers tend to be assigned higher-achieving students, while novice teachers more often teach classrooms with greater concentrations of low-income or minority students (Lankford et al., 2002; Borman and Kimball, 2005; Clotfelter et al., 2006; Kalogrides and Loeb, 2013). Sorting can occur through administrator assignment, parent advocacy, or teacher mobility across schools (Boyd et al., 2011; Goldhaber et al., 2015; Paufler and Amrein-Beardsley, 2014). Such sorting could explain the link between classroom characteristics and evaluation scores.

A second explanation, however, is that classroom context has a causal influence on ratings. This could occur if teaching is genuinely more difficult in more challenging settings, thereby depressing a teacher's observable performance. New teachers often work both in disorganized schools and in a less supportive environment than other teachers, which may further depress ratings (Kraft and Papay, 2014). Alternatively, it could reflect rater bias, where evaluators subconsciously penalize or reward teachers based on the students they teach, independent of the teacher's actual effectiveness. Distinguishing between sorting and these causal channels is critical for assessing the validity and fairness of observation-based evaluations.

To isolate the causal effect of classroom composition, a set of studies take advantage of the experimental design of the MET project, which randomly assigned students to teachers within schools (Steinberg and Garrett, 2016; Campbell and Ronfeldt, 2018; Cherng et al., 2022). This design mitigates concerns about within-school teacher sorting. These studies find that classroom context matters: Steinberg and Garrett (2016) and Campbell and Ronfeldt (2018) both find that teachers receive higher observation ratings when assigned higher-achieving students. The latter authors believe both rater bias and actual differences in instructional quality may explain their estimates, but

cannot disentangle the effects of the two. Cherng et al. (2022) finds that teachers in classrooms with higher proportions of Black and Latinx students tend to receive lower ratings, regardless of the teacher's own race. In our analysis, classroom demographic composition is not systematically related to overall observation scores after controlling for students' baseline achievement and behaviors; however, a higher share of Black students is associated with lower ratings in the classroom environment and instruction domains. While compelling, a limitation of the MET studies is that ratings were conducted by trained external observers on videotaped lessons for a small sample of teachers, a context that differs from the high-stakes, in-person evaluations conducted by school principals that are the norm in practice.

This study builds on and extends this literature in several ways. First, we analyze authentic, high-stakes evaluations conducted by school administrators in CPS. Our data include evaluation ratings for the universe of teachers in CPS, the then third largest school district in the US, and include teachers in general education, math and English subjects as well as arts, music, and other subjects. Therefore, the findings in this paper are more generalizable for policy and practice. Second, we employ a teacher-by-school fixed-effects design that leverages naturally occurring year-to-year variation in classroom composition for the same teacher, and at the same time control for time-varying teacher characteristics, such as teaching experience. Under the identifying assumption of exogenous classroom shocks, this quasi-experimental design identifies the causal effects of classroom context while mitigating concerns about the sorting of teachers to schools. Third, we contrast the effects on high-stakes administrator ratings with those on low-stakes student surveys, providing insight into whether these two perspectives are differentially sensitive to classroom context. Finally, we simulate a policy adjustment to ratings, quantifying the potential equity implications of accounting for classroom composition in teacher evaluations.

3 Teacher Evaluation in Chicago Public Schools

In response to state legislation, Chicago Public Schools (CPS) implemented a new teacher evaluation system, *Recognizing Educators Advancing Chicago* (REACH), beginning with non-tenured teachers in 2012–13 and expanding to all teachers in 2013–14.[1] REACH was designed to improve instruction

[1]See Sartain et al. (2020) for a detailed description of REACH.

and student learning by providing a clear definition of high-quality teaching and supporting ongoing professional growth (Chicago Public Schools, 2019). Teachers receive a composite score on a 100–400 scale, which is then converted into one of four categories: *Unsatisfactory*, *Basic*, *Proficient*, or *Distinguished*. Despite this formative intent, the ratings carry significant stakes. Tenured teachers with unsatisfactory or basic ratings face remediation plans and potential dismissal, while non-tenured teachers with these ratings cannot progress toward tenure and are subject to dismissal.

REACH evaluation scores are based on two primary components: classroom observation ratings (70 percent of the final score) and student growth metrics (30 percent of the final score). Student surveys are not part of teacher evaluations.

3.1 High-stakes teacher performance measures

Classroom observations. Each teacher is observed four times during the evaluation cycle by a certified principal or assistant principal, in both announced and unannounced visits. CPS uses an adaptation of the Charlotte Danielson Framework for Teaching, which structures teacher practice into four domains: Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibility. These domains are further broken down into 19 specific components (e.g., Managing Student Behavior, Using Assessment in Instruction). Each component is rated on a four-point scale, and these ratings are aggregated into a single teacher practice score. Appendix Table A.1 provides a list of the domains and components, while Appendix Table A.2 shows a sample rubric.

Student growth. This metric is calculated using one of two methods. The first method uses value-added measures (VAMs) based on the Northwest Evaluation Association (NWEA) test for grades 2–8. These measures are used for teachers in tested subjects in reading and math. The second method, which applies to all other teachers, is based on performance tasks. These are subject- and grade-level-specific assessments developed by the district and administered by teachers at the beginning and end of the year. Student growth is then calculated based on the scores from these tasks. There is already an extensive body of literature on the construction and use of VAMs, which is not the focus of this paper.

Because most teachers do not teach tested grades or subjects, classroom observations determine the majority of REACH ratings in practice. Even for teachers with value-added scores, observations

carry the heaviest weight. Teachers also tend to value the feedback from these observations, making it especially important to understand whether scores reflect instructional quality or classroom composition.

3.2 Low-stakes student surveys

In addition to the formal evaluation system, CPS administers an annual, low-stakes survey to students in grades 6–12 through a partnership with the University of Chicago Consortium on School Research. This 5Essential survey captures various aspects of school climate and culture. With student response rates around 80 percent, improvements on these constructs are associated with school-level achievement gains (Bryk et al., 2010; Hart et al., 2020). While the survey results are publicly available and included in the school accountability system, they do not directly influence individual teacher evaluations, distinguishing them from the REACH system.

A key component of the survey asks students to report on their experiences in a specific, randomly selected class (English, math, or science). These course-specific items cover topics such as teacher expectations, coursework relevance, and classroom behavior. We use these granular, course-specific items because they can be linked to teacher-level administrative data, providing a unique measure of teacher practice from the student perspective. Appendix Table A.3 lists the specific survey items and their associated indices.

4 Methodology and Data

4.1 Conceptual Framework

We define evaluator bias as a systematic deviation in teacher ratings that is attributable to classroom characteristics, independent of the teacher's true quality (Grissom and Bartanen, 2022). Formally, let q_{it} represent the latent quality of teacher i in year t. While unobserved, this quality is reflected in teaching practices that evaluators rate to produce an overall score, Y_{it}. We model this relationship as:

$$Y_{it} = q_{it} + \delta C_{it} + \tilde{\varepsilon}_{it} \tag{1}$$

where C_{it} is a vector of classroom characteristics (e.g., baseline achievement and behaviors) and $\tilde{\varepsilon}_{it}$ is an idiosyncratic error term. The parameter of interest, δ, captures the magnitude and direction of evaluator bias. A non-zero δ implies that classroom characteristics directly influence ratings, holding teacher quality constant.

Observational data often show a correlation between student characteristics and teacher ratings. Figure 1 illustrates this pattern in our data, plotting the distribution of teacher observation scores (Panel A) and survey scores (Panel C) conditional on the classroom's average baseline student test scores. As shown in Panel A, teachers whose classrooms are in the bottom quartile of prior student achievement are disproportionately rated in the bottom quartile of observation scores (40 percent) and rarely in the top quartile (11 percent). The reverse is true for teachers with students in the top achievement quartile. This correlation could reflect true causal bias ($\delta \neq 0$) or, alternatively, the non-random sorting of teachers to classrooms that would produce a spurious correlation between q_{it} and C_{it} (e.g., highly effective teachers sorting to more advantaged schools and classrooms).

The ideal experiment to disentangle these channels would randomly assign teachers to higher-achieving and lower-achieving classrooms, so that teachers are balanced across settings, and test for systematic differences in teacher performance ratings. In the absence of such experiment, we employ a quasi-experimental approach.

We decompose teacher quality into a time-invariant component, q_i, and a transitory shock, θ_{it}, such that $q_{it} = q_i + \theta_{it}$. Substituting this into Equation 1 yields our estimating equation:

$$Y_{it} = q_i + \delta C_{it} + \varepsilon_{it} \tag{2}$$

where the composite error term is $\varepsilon_{it} = \tilde{\varepsilon}_{it} + \theta_{it}$. Intuitively, if we compare the same teacher across years—holding the teacher's time-invariant attributes fixed—any systematic association between year-to-year changes in classroom composition and year-to-year changes in ratings reflects δ, provided within-teacher shifts in composition are as-good-as random with respect to time-varying shocks to quality. This strategy requires the following identification assumption:

Assumption 1 (Exogeneity of classroom shocks) *Changes in a teacher's classroom composition are uncorrelated with contemporaneous shocks to teacher quality, conditional on time-invariant*

teacher characteristics:

$$\mathbb{E}[\varepsilon_{it}|C_{it}, q_i] = \mathbb{E}[\varepsilon_{it}|q_i] = 0 \tag{3}$$

This assumption allows for systematic sorting, such as more effective teachers consistently being assigned to higher-achieving students. It would be violated, however, if teachers who experience improvements in their effectiveness are systematically assigned different types of classrooms in subsequent years. One potential threat therefore arises from the joint dynamics of teacher experience and student assignment. The literature documents that teacher effectiveness grows with experience and that more experienced teachers are often sorted into different classrooms (Kraft and Papay, 2014; Kalogrides and Loeb, 2013). To mitigate this concern, our main specifications control for a flexible function of teaching experience. We further assess the plausibility of Assumption 1 through a series of robustness and placebo tests in Sections 5.4 and 5.5.

4.2 Data and key variables

We use de-identified administrative and survey data from CPS for the 2012–13 to 2016–17 school years. These longitudinal data include student demographics, test scores, attendance, 5Essential surveys, teacher personnel files, and REACH evaluation data.

Teacher performance ratings. Our primary dependent variables are two measures of teacher performance.

- Classroom observation score: The final score from the district's teacher evaluation system, averaged across four domains: *planning and preparation*, *classroom environment*, *instruction*, and *professional responsibilities*. Scores for each domain serve as secondary outcomes. Virtually all teachers (98 percent) have complete domain data.

- Survey score: The 5Essentials student survey includes 36 course-specific items for a randomly selected class (English, math, or science). Items are grouped into seven theory-based indices: *peer group for academic work*, *classroom rigor*, *academic press*, *course clarity*, *academic engagement*, *academic personalism*, and *classroom disruption*. We compute student-level means for each index, aggregate to teacher-year means, and average indices to a single teacher-level survey score.[2] The individual indices are secondary outcomes.

[2]Items are on a 1–4 Likert scale, and those originally on 1–5 are mapped to 1–4. Three items are reverse-coded

Classroom characteristics. We construct a rich set of classroom-level characteristics by aggregating student-level data. Student characteristics include demographics (gender, race/ethnicity, free or reduced-price lunch FRPL status, special education status), behaviors (lagged attendance, whether suspended, and whether repeating the grade), and prior academic achievement (lagged GPA and standardized test scores). Test scores are from the NWEA exam, and we average math and reading scores and normalize them by grade and year. We use lagged achievement and behavior variables to mitigate simultaneity bias, as these could be influenced by the teacher's current-year effectiveness. We also include class size.

Classroom quality index. To reduce the dimensionality of multiple classroom characteristics, we construct a summary measure of classroom context using principal component analysis (PCA). The classroom quality index is defined as the first principal component of student-level baseline achievement and behavioral measures: standardized test scores, GPA, attendance, suspensions, and grade repetition status. We aggregate this index to the classroom level. A higher value indicates a classroom that, on average, consists of higher-achieving students with fewer behavioral issues (see Appendix Table A.4 for component loadings). The first principal component explains 40 percent of the joint variation in these underlying measures.[3] Finally, we note that the composite index also simplifies downstream analyses—such as heterogeneity and policy simulations—by allowing us to study interactions with a single, interpretable measure of classroom context rather than a high-dimensional vector of correlated variables.

Teacher characteristics. We control for teacher gender, race/ethnicity, tenure status, age, years of experience within the district (years since CPS hire), educational level, and teaching credentials. These variables are derived from district personnel files.

so higher is better. Missing values are ignored in the calculation of the index averages.

[3]Focusing on the first component may exclude some variation in the underlying measures. To assess the sensitivity of our results to this choice, we present specifications that additionally control for subsequent principal components (Appendix Table A.6) as well as specifications that include the full vector of baseline achievement and behavioral characteristics directly (Appendix Table A.7). In all cases, our main results remain qualitatively similar. We therefore focus on the first component for parsimony and ease of interpretation.

4.3 Sample description

We use student transcript files to link students to teachers and construct classroom rosters.[4] All students taught by the same teacher in a year are pooled into a single roster, as the administrative data do not record actual classroom assignments and REACH data do not identify which section was observed. Pooling classrooms avoids potential selection bias if a teacher with multiple sections were more likely to be observed in the most advantaged class.[5]

Our sample includes all grades 3–8 teachers with rosters of at least five students with lagged test scores. We exclude classrooms with more than 25 percent special education students, since co-teaching arrangements may affect ratings.[6] After applying these restrictions, 94 percent of teachers match to REACH evaluation data and 36 percent match to student survey records (reflecting eligibility rules for middle-grade English, math, and science). Given the high matching rate to the REACH data, we restrict the sample to these teachers.

Our analysis uses two primary samples. The classroom observation sample includes 30,479 teacher-year observations from 10,934 unique teachers. The survey sample includes 10,991 observations from 3,345 unique teachers. Nearly all eligible teachers ($\approx$ 97 percent) have survey data, and student response rates are about 90 percent among eligible students. Teachers appear in the data for about three years on average.

Table 1 presents descriptive statistics. In the observation sample, the typical teacher is female (83%), White (47%) or Black (26%), around 40 years old, with nine years of CPS experience, and 74% are tenured. Nearly all hold a college degree (96%) and teaching certification (98%), about 70% hold a master's degree, and 8% hold National Board Certification. Teachers are linked to an average of 62 students. Classrooms are on average 38% Black, 46% Hispanic, and 11% White, and

[4]The transcript files list all courses taken by each student and have identifier for the teacher who assigned their final grades. These files also allow us to identify each student's English, math, and science teachers and link them to student survey responses.

[5]A limitation of our linking procedure is that the teacher who assigned the final grade may not be the student's regular teacher. Although we cannot verify it, we believe that this is not very common across schools. Furthermore, if a school has a dedicated person who submits grades, she would be dropped from the analytic sample (as we describe below) if the number of linked students crosses a threshold. Another limitation is measurement error in class characteristics for teachers with multiple classes. Because we do not observe which classroom the observation occurred, the characteristics of our constructed roster may not correspond to the actual classroom. To address this limitation, we restrict the sample to teachers who are likely to have a single classroom as a robustness check.

[6]We perform several sample restrictions: Starting with grades 1–8 teachers, we first exclude rosters with less than five students with lagged test scores to reduce imprecision when computing classroom mean test scores (17%). Second, we exclude rosters with more than 25 percent of special education students (15%). Third, we drop rosters with more than 200 students because such students are likely to be mislinked (21%).

predominantly low-income (85% FRPL); however, classrooms are highly racially segregated (see Figure 2 Panel C). Classroom observation scores cluster between 3 and 4, while student survey scores concentrate near 3 on the 1–4 scale (see Figure 2 Panels A and B).

Appendix Table A.5 shows that classroom observation and student survey scores are moderately correlated with each other (ρ=0.24). Both are also positively correlated with teacher value-added measures (ρ=0.22 for observation, and ρ=0.17 for surveys). These modest correlations align with prior research (e.g., Kane and Staiger, 2012; McCaffrey et al., 2013; Kane et al., 2011) and suggest that the measures capture complementary but distinct dimensions of teacher effectiveness.

4.4 Empirical strategy

We estimate the causal effect of classroom characteristics on teacher ratings using a teacher-by-school fixed effects model. This quasi-experimental design leverages variation in the composition of students assigned to the same teacher in the same school over different years. Our main specification is:

$$Y_{ist} = \delta C_{ist} + \beta X_{it} + \gamma_{is} + \tau_t + \varepsilon_{ist} \tag{4}$$

where Y_{ist} is the performance rating (e.g., observation score) for teacher i in school s in year t. C_{ist} is the vector of classroom characteristics, with δ being the vector of coefficients of interest. X_{it} is a vector of teacher characteristics. Classroom characteristics include the classroom quality index, share of male students, the shares of Black and non-Black non-Hispanic students, with Hispanics as the omitted racial group for being the largest group, share of FRPL students, and share of special education students, noting that classrooms with large proportion of these students were excluded. We also control for log of roster size. Teacher characteristics include teacher gender indicator, race and ethnicity indicators, age, experience and experience squared, tenure status indicator, indicator for not having a college degree and another indicator for not having any teaching certificates given that most teachers have a college degree and teaching certification, indicator for having a graduate degree, and an indicator for obtaining the National Board Certification.[7]

The model includes teacher-by-school fixed effects (γ_{is}) and year fixed effects (τ_t). The γ_{is}

[7]Missingness in teacher covariates is rare (less than 2 percent). We impute means for continuous variables (e.g., age) and zeros for indicators, and include missing-value flags in regression models. For experience missing at entry, we set experience to 0 in the first observed year and increment thereafter.

terms absorb all time-invariant teacher characteristics (e.g., race/ethnicity, teaching style, course difficulty) and any stable teacher sorting patterns into specific schools and any systematic student sorting into specific teachers and courses. The τ_t terms control for district-wide shocks common to all teachers in a given year. Robust standard errors are reported, except for specifications without fixed effects, in which standard errors are clustered at the teacher level.

The identifying variation for δ comes from year-to-year changes in the types of students a teacher teaches. Figure 3 shows that this variation is significant. For example, the total standard deviation for our classroom quality index is 0.82 SD and the within-teacher variation is 0.31 SD, representing 14 percent ($= 0.31^2/0.82^2$) of the total variation. Similarly, the within-teacher variation for classroom observation and survey scores is 19 and 36 percent of their respective total variation. For classroom demographics, the share of variation attributable to within-teacher changes ranges from as little as 0.8 percent (share Black students) to as much as 60 percent (share male students). These figures confirm we have sufficient variation to identify the effects of interest.

5 Results

5.1 Effects of classroom characteristics on evaluator ratings

We begin by examining the effect of classroom characteristics on teachers' observation scores using the fixed-effects model specified in Eq. 4. For ease of interpretation, both the dependent variables and the classroom quality index are standardized by year to have a mean of zero and a standard deviation of one.

Table 2 presents the main results. Columns 1–3 progressively introduce controls. In a specification with only classroom demographics and year fixed effects (Column 1), a one standard deviation increase in the classroom quality index is positively and significantly associated with 0.17 SD higher observation scores. Several demographic characteristics also show significant correlations; for instance, a 10 percentage point increase in the share of Black students is associated with a 0.05 SD decrease in scores, while a similar increase in non-Black, non-Hispanic students is associated with a 0.02 SD increase. These relationships persist after controlling for teacher characteristics (Column 2) and school fixed effects (Column 3), indicating that even within the same school, teachers assigned to classrooms with higher-achieving students and fewer underserved students tend to receive higher

ratings.

Column 4 shows our preferred specification, which includes teacher-by-school fixed effects to isolate the effect of year-to-year changes in a teacher's classroom assignment within a given school. In this specification, a 1 SD increase in the classroom quality index is associated with a 0.068 SD increase in observation scores. On the other hand, demographic coefficients become statistically indistinguishable from zero; the one exception is class size, which becomes negatively significant, though the magnitude is small (a 10 percent increase in class size is associated with a 0.004 SD reduction in ratings). We note the large increase in the R-squared from 0.20 to 0.80 once teacher-by-school fixed effects are included; it is consistent with substantial persistent, time-invariant differences across teachers explaining most variation in ratings.

To probe the mechanism, Columns 5 and 6 estimate the models with the classroom quality index alone and demographics alone while retaining the teacher-by-school fixed effects. The classroom quality effect remains stable at 0.077 SD per 1 SD increase. When the quality index is omitted, some demographic variables appear significant; once the index is included, those effects largely attenuate. This pattern suggests that baseline achievement and behavior, rather than demographics per se, drives most of the link between classroom composition and observation ratings.[8]

We confirm this interpretation in Appendix Table A.7, which replaces the classroom quality index with the full vector of student achievement and behavioral characteristics. Without teacher-by-school fixed effects, nearly all characteristics are statistically significant (Column 1). With teacher-by-school fixed effects, however, only baseline test scores, GPA, and attendance rate remain significant predictors (Column 2), while most demographic variables do not. In terms of magnitude, a 1 SD increase in average baseline test scores and GPA is associated with a 0.04 and 0.03 SD increase in classroom observation scores, respectively. Similarly, a 1 SD increase in prior-year average attendance rates is associated with a 0.02 SD increase in observation scores. Other behavioral measures are less precisely estimated: the share of students who were suspended is negatively related to observation scores, while the share of repeaters is positively related, though

[8]While the first principal component captures the dominant common variation across achievement and behavioral measures, subsequent components also contain information. Appendix Table A.6 shows that the second principal component is also a statistically significant predictor of teacher performance ratings, reinforcing that meaningful information is captured beyond the first component. Importantly, however, including these additional components does not alter our main conclusions regarding the relationship between classroom context and subjective teacher performance ratings.

both coefficients are imprecise. In contrast, classroom demographic composition—including gender, race/ethnicity, socioeconomic status, and special education status—has effects that are statistically indistinguishable from zero at conventional levels, with the exception of class size, which has a small negative effect. Overall, these results indicate that prior achievement and attendance are the primary drivers of the classroom composition effect, supporting our use of a composite index to summarize these dimensions and reduce dimensionality in the main analysis.

Do these effects operate across all domains of the observation rubric? We disaggregate the analysis by the four domains that comprise the observation score in Appendix Table A.8. The classroom quality index is a consistent and positive predictor of ratings across all domains. The effects are largest for in-classroom domains—Classroom Environment (0.09 SD) and Instruction (0.07 SD)—but also spill over to out-of-classroom domains like Planning and Preparation (0.05 SD) and Professional Responsibilities (0.05 SD). These spillovers may arise if a single evaluator's bias persists across all domains or if genuine teacher performance improvements, boosted by a higher-achieving class, extend beyond direct instruction. Consistent with the main results, most demographic characteristics are not robust predictors of domain scores; one exception is that higher Black student share is associated with slightly lower Classroom Environment (-0.03 SD per 10 p.p.) and Instruction (-0.02 SD per 10 p.p.) ratings.

5.2 Effects of classroom characteristics on student survey reports

We next examine the influence of classroom context on student survey reports. Because surveys are fielded in grades 6–12 for middle grade English, math, and science teachers, we first replicate the observation score analysis on the survey sample (Appendix Table A.9). For this sample, the classroom quality effect on observation scores is larger. Our preferred specification implies that a 1 SD increase in classroom quality raises observation ratings by about 0.13 SD, and classroom demographics are not statistically significant.

Table 2 (Columns 7–12) presents the main results for student survey scores. In our preferred specification with teacher-by-school fixed effects (Column 10), a 1 SD increase in the classroom quality index is associated with a 0.15 SD increase in survey scores. This effect is robust to the inclusion or exclusion of demographic controls and is similar in magnitude to the effect on observation scores for this same middle-school subsample, suggesting that teacher performance ratings are more

sensitive to classroom quality in middle grades than in elementary grades.

Contrary to the findings for observation scores, some demographic characteristics remain significant predictors of student survey reports even after controlling for the classroom quality index. Specifically, a 10 percentage point increase in the share of Black students is associated with a 0.09 SD decrease in survey scores, while a similar increase in the share of FRPL-eligible students is associated with a 0.09 SD increase. Larger classes are associated with lower ratings, with the magnitude of this effect being larger for surveys than for observations. Appendix Table A.7 (Columns 5–8), which replaces the classroom quality index with the full vector of student achievement and behavioral characteristics, indicates that average baseline test scores and the share of students suspended in the prior year appear to drive much of the classroom quality index effect on survey outcomes.

Do these patterns hold across survey domains? Appendix Table A.10 disaggregates these findings by the seven domains of the student survey. The classroom quality index is positively associated with scores on all domains except for Academic Engagement, whose impact is positive but imprecise. Demographics also show domain-specific effects; for example, the share of Black students is negatively associated with several domains, while the share of FRPL-eligible students is positively associated with all domains. Class size negatively affects all domains. Notably, student reports on Classroom Disruptions are sensitive to a wide range of classroom characteristics. These results suggest that student perceptions are not immune to idiosyncratic changes in classroom composition.

5.3 Heterogeneity of classroom effects by teacher characteristics

To explore whether these effects are stable across different types of teachers, we test for heterogeneity by teacher demographics and prior effectiveness. All models use our preferred specification (classroom demographics, teacher covariates, and teacher-by-school fixed effects).

Table 3 reports results separately by teacher gender and race/ethnicity (Columns 1–6). The positive effect of the classroom quality index on observation scores (Panel A) is present and statistically significant for most subgroups, with point estimates ranging from 0.07 to 0.10 SD. For Hispanic teachers, the effect is smaller and imprecise; however, we cannot reject equality of effects across groups. For student survey scores (Panel B), the effect of classroom quality is also consistently positive (0.15–0.22 SD), though the estimate for Black teachers is smaller and not statistically significant.

Table 4 presents heterogeneity results by baseline teacher quality, measured by either prior-year observation score (Columns 1–3) or value-added (Columns 4–6). Panel A Columns 1 and 4 confirm the main findings for the subset of teachers with available prior-year data: a 1 SD increase in classroom quality raises observation scores by about 0.05–0.06 SD, and this estimate is stable when controlling for baseline quality (Columns 2 and 5). To test whether higher-quality teachers benefit more from higher-achieving classrooms, we interact the classroom quality index with the baseline quality measures (Columns 3 and 6). For observation scores, the interaction with prior observation score is marginally significant ($p<0.10$), but the interaction with prior value-added is not. For survey scores (Panel B), neither interaction term is statistically significant. Overall, we find little evidence that the effect of classroom quality differs systematically by teachers' demographic characteristics or their prior performance. The positive classroom quality effect is broadly similar across teacher types.

5.4 Robustness checks

We conduct two main robustness checks to address potential threats to validity. First, a potential concern is dynamic sorting, whereby our results are driven by a spurious correlation between classroom assignments and unobserved, time-varying teacher skill. For example, teachers who are improving might be systematically assigned higher-achieving students. Our main specification already controls for experience (and experience squared), and Table 4 shows the results are robust to controlling for prior-year effectiveness. To further address this, we restrict the sample to experienced teachers (5+ years), whose effectiveness is likely more stable (Rockoff, 2004). As shown in Table 4 Column 7, the classroom quality effect remains similar in magnitude for this group for both observation and survey scores. This suggests our findings are not primarily driven by the sorting of teachers on unobserved gains in effectiveness.

Second, for teachers with multiple class sections, our pooled classroom roster may introduce measurement error. The characteristics of the pooled roster may not perfectly match those of the specific classroom section that was observed. To mitigate this, we restrict the analysis to two subsamples where measurement error is less likely: teachers with small class sizes ($\leq$35 students) and elementary school teachers (grades 5 and below), who typically have self-contained classrooms. The results for these subsamples (Table 4 Columns 8 and 9) show statistically significant effects of

similar magnitude (0.05–0.07 SD), suggesting that such measurement error is not a major driver of our results. For completeness, we estimate effects for middle school teachers (Column 10) and find a larger coefficient (0.125 SD), mirroring the magnitude found for student survey scores, which are from middle grades.

5.5 Plausibility of exogenous variation

We assess Assumption 1, that within-teacher changes in classroom composition are as good as random with respect to time-varying determinants of teacher quality, using placebo tests. If classroom shocks are exogenous, the classroom quality index should not predict past or future performance once teacher fixed effects are included.

Table 5 reports regressions of teacher performance on the current classroom quality index, replacing the outcome with lagged ($t-3$, $t-2$, $t-1$) and lead ($t+1$, $t+2$, $t+3$) observation and survey scores (Panels A and B). Across both outcomes, the coefficient on classroom quality is statistically indistinguishable from zero in all lag and lead years (Columns 1–3 and 5–7), with the exception of a single marginally significant estimate for observation scores at $t+1$. Given the number of placebo tests conducted, this isolated finding should be interpreted with caution. In contrast, the effect in the concurrent year (t) is positive and highly statistically significant (Column 4). Overall, this pattern—where year-to-year shifts in classroom composition do not systematically predict prior or subsequent ratings conditional on teacher-by-school fixed effects—is consistent with the identifying assumption that the remaining within-teacher variation in classroom composition is orthogonal to time-varying shocks to teacher effectiveness.

We further test for correlated shocks by examining whether changes in the classroom quality index are associated with changes in other classroom demographics (Appendix Table A.11). Without fixed effects, demographic shares (e.g., male, Black, FRPL, special education) strongly predict lagged, current, and future classroom quality (Columns 1, 3, and 5). With teacher fixed effects (Columns 2, 4, and 6), contemporaneous correlations remain as one would expect—higher-achieving classes are, in the same year, associated with fewer male, Black, FRPL, and special education students and more non-Black non-Hispanic students (Column 4)—but these relationships largely disappear in lagged and future years, with the exception of special education at $t+1$. The absence of predictive power outside the current year supports the assumption that the identifying variation

arises from idiosyncratic within-teacher shocks rather than persistent multi-year sorting trends. Together, the placebo and correlated shock tests increase our confidence that the estimated effects capture causal impacts of classroom composition on performance ratings.

5.6 Exploring mechanisms: student-teacher interaction vs. evaluator bias

Having established that classroom composition influences teacher ratings, we now explore the underlying mechanisms. Classroom effects on teacher performance ratings could generate from at least two channels. First, evaluator bias: raters may (consciously or not) attribute favorable classroom behaviors or achievement to the teacher, inflating scores when students are easier to teach. Second, student-teacher interaction: classroom composition may influence actual teaching practice (e.g., pacing, classroom management, differentiation), thereby raising genuine performance.[9] Both mechanisms are consistent with prior work documenting sensitivity of classroom observation scores to incoming achievement and racial/ethnic composition (e.g., Campbell and Ronfeldt 2018; Steinberg and Garrett 2016; Whitehurst et al. 2014) and with evidence of disparate teacher impacts by student characteristics and student-teacher demographic match effects (e.g., Loeb et al. 2014; Aucejo et al. 2022; Biasi et al. 2021; Gershenson et al. 2022; Egalite et al. 2015; Dee 2004, 2007; Delgado 2025).

In our conceptual model, consider observed teacher practice p_{it} to depend on teacher quality q_i and classroom quality C_{it}:

$$Y_{it} = p_{it}(q_i, C_{it}) + \varepsilon_{it} \tag{5}$$

If C_{it} shifts p_{it} (student-teacher interaction), we should see classroom quality matter for outcome-based productivity measures as well, not only for subjective ratings. If, instead, the effect operates mainly through evaluator bias, then classroom quality will shift subjective ratings but have little systematic effect on outcome-based measures.

We implement two empirical tests. First, heterogeneity by baseline teacher quality. If student-teacher interaction (i.e., match) is the primary channel, higher-quality teachers might benefit differentially when assigned higher-achieving classes. As mentioned before, in Table 4, the interaction

[9] Certainly, there are other potential sources of bias. For example, an often-mentioned potential concern is that the classroom observation instrument itself is subject to cultural biases, as it may center Euro-centric classroom practices that are not appropriate to use with all student populations. If this is the case, we might expect teachers in classrooms with high shares of White students to receive higher ratings than teachers of students of color because of bias in the instrument and not because of differences in actual teacher quality.

terms between classroom quality and prior-year teacher effectiveness (measured by either observation scores or value-added) are consistently small and statistically indistinguishable from zero for both observation and survey scores. This suggests that teachers of all baseline effectiveness levels benefit similarly from being assigned a higher-achieving class, consistent with evaluator bias or with a broadly similar interaction effect across teachers rather than large, systematic amplification for higher-quality teachers.

Second, we examine teacher productivity measures based on student learning. If higher-achieving classrooms genuinely make teachers more effective, this enhanced productivity should translate into greater student test score growth. Conversely, if the effect on subjective ratings is driven by evaluator bias, there should be no corresponding positive effect on student-outcome-based measures.

We test this using two student-growth metrics: district-developed performance tasks (all subjects) and value-added (standardized English and math tests). Table 6 presents the results. While a 1 SD increase in the classroom quality index raises observation scores by 0.07–0.08 SD (Columns 1 and 4), the same change is associated with a statistically significant decrease in both performance task scores (Column 3) and value-added scores (Column 6). The negative relationship may arise because students with higher baseline achievement have less room to grow. Regardless of the reason, the key finding is the stark opposition: classroom characteristics that positively predict subjective performance ratings negatively predict objective measures of student learning. This result is difficult to reconcile with the student-teacher interaction hypothesis and points instead toward evaluator bias as the primary mechanism.

In summary, our quasi-experimental results show that both evaluator and student ratings are positively influenced by the composition of the classroom, particularly the prior achievement and behavior of students. Our exogeneity tests suggest this relationship is plausibly causal. Furthermore, our exploration of mechanisms indicates that these effects are not reflected in outcome-based measures of teacher productivity, pointing to evaluator bias rather than genuine changes in teacher effectiveness as the likely driver. Given these findings, by accounting for the influence of classroom composition, educational systems can develop fairer and more accurate evaluation systems, better supporting teacher development and improving student outcomes. Strategies could include training evaluators, benchmarking teachers against those with similar classrooms, or applying statistical adjustments to performance ratings.

6 Policy Simulations

Classroom observation scores are a critical component of teacher evaluations, often carrying the most weight in personnel decisions. For example, in Washington, D.C. Public Schools (DCPS), observation scores account for 40%–65% of a teacher's evaluation, while student surveys are weighted at 10% (District of Columbia Public Schools, 2022). In our context, observation scores account for a substantial 70% of the overall evaluation, and student surveys are currently excluded. Given our evidence that teachers assigned to classrooms with more disadvantaged students tend to receive lower observation scores, this section conducts a policy counterfactual to assess the impact of adjusting performance ratings for classroom composition.[10] We identify which teacher groups would benefit from a more equitable evaluation system.

6.1 Adjusting teacher performance ratings for classroom composition

We investigate two approaches to mitigate the influence of classroom context on subjective performance measures.

Tournament-style adjustment. The first approach compares teachers only within groups of classrooms that share a similar composition, akin to establishing a tournament within more homogeneous contexts. Teachers are divided into two (below- and above-median) or three (low-, medium-, and high-index) groups based on the classroom quality index. A teacher's adjusted rank is then determined by their performance within their composition-based peer group. This approach directly addresses sorting by comparing like-with-like.

Regression-based adjustment. The second approach is a regression-based adjustment similar to methods used for value-added modeling (VAM). Just as VAM adjusts student test scores for prior achievement and student demographics to isolate the teacher effect, we adjust observation scores by classroom quality index, thus increasing scores for teachers with more disadvantaged students and decreasing scores for those with more advantaged students.

We first estimate the influence of classroom characteristics on performance ratings by modifying

[10] Most variation in classroom quality occurs between schools rather than within schools. In Appendix Table A.12, classroom quality is regressed on teacher characteristics with and without school fixed effects. Across schools, Black and Hispanic teachers are assigned classrooms with, on average, 0.9 and 0.2 SD lower achievement and behavioral indices, respectively, than White teachers. Within schools, these gaps shrink to about 0.06 SD, suggesting that differences in classroom composition largely reflect cross-school sorting. The counterfactual policies we simulate are thus designed to mitigate inequities arising from such sorting.

Equation 4:

$$Y_{ist} = \alpha + \delta C_{ist} + \tau_t + \varepsilon_{ist} \tag{6}$$

where Y_{ist} is the performance rating (observation or survey score) for teacher i in school s at time t, C_{ist} is the classroom quality index, and τ_t is year fixed effects. This model excludes teacher characteristics and teacher-by-school fixed effects. We then define the adjusted score as the residual from this regression:

$$e_{ist} = Y_{ist} - \hat{Y}_{ist} = Y_{ist} - (\hat{\alpha} + \hat{\delta} C_{ist} + \hat{\tau}_t) \tag{7}$$

This residual e_{ist} represents the portion of the performance rating that is unexplained by the classroom composition and is thus our quality-adjusted performance score.

Figure 1 compares the distributions of unadjusted and adjusted scores. For classroom observation scores, the adjustment attenuates the relationship between test score quartiles and observation ratings (Panels A and B). After adjustment, teachers are more evenly distributed across observation score quartiles regardless of their students' baseline achievement. For student surveys, the distribution remains largely unchanged (Panels C and D), as the unadjusted distribution was evenly distributed.

Properties of regression-adjusted scores. The regression-based adjustment preserves the core properties of the ratings. The correlation between adjusted and unadjusted classroom observation scores is high (ρ=0.91). The reliability (year-to-year correlation) slightly decreases for adjusted observation scores (0.69 vs. 0.75). Importantly, the relationship with the district's value-added measure remains nearly unchanged: the rank-rank correlation between teacher VAM and unadjusted observation score is 0.23, compared to 0.21 for the quality-adjusted score (Appendix Figure A.1). Similar stability appears for survey measures: adjusted and unadjusted scores correlate at 0.96, with year-to-year reliabilities of 0.43 and 0.45, respectively, and rank-rank correlations with value-added of 0.19. Overall, adjusting for classroom composition does not substantially change the statistical reliability or alignment of teacher performance metrics.

6.2 Who would benefit from adjusting for classroom composition?

To evaluate the policy impact of adjusting scores, we rank all teachers based on their (1) unadjusted rating (baseline) and (2) adjusted rating (policy counterfactual). The change in ranking is calculated

as the difference between the adjusted and unadjusted percentile ranks (a positive change indicates improvement). We also define two tail indicators: exiting the bottom 5%, an indicator equal to 1 if a teacher moves above the 5th percentile after adjustment, conditional on being below it before, and exiting the top 5% , an indicator equal to 1 if a teacher moves below the 95th percentile after adjustment, conditional on being above it before. We then regress these outcome variables on teacher characteristics.

Within-group tournament policy. Columns 1–3 of Table 7 report the results for the two-group tournament (below vs. above-median classroom quality index). Black teachers experience an average ranking gain of 7.2 percentile points relative to White teachers, while Hispanic teachers gain about 2.1 points. Novice teachers also see a modest improvement (0.7 percentile points), as do male teachers (0.5) and teachers without National Board Certification (0.7). At the distribution's lower tail, Black teachers are 7 percentage points more likely to move out of the bottom 5 percent after adjustment. The three-group tournament (low-, medium-, high-quality classrooms; Columns 4–6) yields qualitatively similar results.

For student surveys, adjusting student survey scores has a negligible impact (Appendix Table A.13). This is expected, as we previously established that survey scores were already equally distributed across classroom settings (Figure 1 Panel C). Black teachers experience a modest 0.2–0.6 percentile improvement, and Hispanic teachers at most 0.2 percentile points.

Regression-based adjustment policy. Columns 7–9 of Table 7 implement the regression-based adjustment (Eq. 6). Under this approach, Black teachers experience the largest gains: an average improvement of 8.3 percentile points, an 11 percentage-point higher probability of exiting the bottom 5 percent, and a 16 percentage-point higher likelihood of remaining in the top 5 percent. Hispanic and novice teachers also benefit, though to a smaller extent.

7 Discussion and Conclusions

This paper provides new evidence on the effect of classroom composition on subjective teacher performance measures, specifically rubric-based classroom observation scores and student survey ratings, using panel data from Chicago Public Schools. School districts rely on these measures to inform high-stakes personnel decisions.

Our findings show that teachers receive significantly higher observation and survey ratings in years when they teach higher-achieving, better-behaved students, even when the same teacher remains in the same school. In contrast, most demographic characteristics, such as student race or low-income status, have null or inconsistent effects. These patterns persist across multiple specifications, placebo tests, and robustness checks, suggesting that the results are not driven by teacher sorting or persistent differences in teacher ability.

Taken together, these findings imply that current evaluation systems may partially conflate teaching effectiveness with the composition of students in a teacher's classroom. Evaluators appear to attribute some portion of student readiness or behavior to teacher skill, potentially rewarding teachers who teach more advantaged students and penalizing those who work with underserved populations. This distortion has important implications for both efficiency and equity in teacher evaluation systems.

From an efficiency standpoint, performance-based personnel decisions may misidentify effective teachers if ratings partly reflect the students they teach rather than their actual practice. Teachers assigned to challenging classrooms could be unfairly labeled as underperforming, discouraging them from remaining in or transferring to schools with higher concentrations of disadvantaged students. Such misclassification could unintentionally exacerbate turnover in precisely the schools most in need of experienced teachers.[11] Nevertheless, even without direct dismissal, such negative evaluations may still carry significant professional and psychological consequences, affecting access to mentoring, development plans, and perceptions of competence.

From an equity standpoint, our policy simulations show that adjusting observation scores for classroom composition would notably improve the relative rankings of Black teachers—by roughly eight percentile points on average—and modestly improve those of Hispanic and novice teachers. Because Black teachers are more likely to teach lower-achieving and higher-need students, unadjusted observation scores systematically disadvantage them. An adjustment for classroom characteristics

[11] In our data, we find insignificant impacts of classroom quality index on teacher turnover. Appendix Table A.14 reports the effect of classroom quality index on the likelihood of exiting the school in the following year. Although teachers assigned to higher-quality classrooms are somewhat less likely to leave (Column 2), this association becomes statistically insignificant once teacher fixed effects are included (Column 3). We also find null associations for the sample of teachers who are at risk of being dismissed, including untenured teachers (Column 4) and teachers who had a low rating in the prior year (Column 5). This suggests that while classroom composition influences observation scores, its effect on more consequential outcomes such as school exit is muted. The weak relationship may reflect the limited formal link between low ratings and dismissal, as teachers typically enter a probationary or remediation phase before termination.

could reduce this inequity, raising fairness in evaluation outcomes without substantially diminishing the reliability or predictive validity of the ratings.

Nevertheless, adjustments must be applied with care. If more effective teachers tend to sort into higher-achieving classrooms, part of the observed classroom-quality premium may reflect genuine differences in performance rather than evaluator bias. In that case, mechanically adjusting scores could over-correct, masking true variation in teaching effectiveness. The appropriate balance depends on the underlying mechanism—evaluator bias versus student-teacher interaction—and on the objectives of the evaluation system.

More broadly, these findings underscore the difficulty of designing evaluation systems that isolate teaching quality from contextual factors. While much of the prior debate has focused on potential bias in value-added or student-growth measures, our results indicate that subjective observation-based measures are also sensitive to classroom composition. Adjusting observation ratings for classroom composition—or designing evaluations that explicitly control for it—can improve fairness, particularly for teachers serving historically underserved students. By ensuring that evaluation systems measure teaching skill rather than classroom assignment, districts can make more equitable and accurate personnel decisions, support teacher development, and ultimately improve outcomes for all students.

References

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics*, 25(1):95–135.

Aucejo, E., Coate, P., Fruehwirth, J. C., Kelly, S., and Mozenter, Z. (2022). Teacher effectiveness and classroom composition: Understanding match effects in the classroom. *The Economic Journal*, 132(648):3047–3064.

Biasi, B., Fu, C., and Stromme, J. (2021). Equilibrium in the market for public school teachers: District wage strategies and teacher comparative advantage. Nber working paper 28530, National Bureau of Economic Research.

Borman, G. D. and Kimball, S. M. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School Journal*, 106(1):3–20.

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., and Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, 48(2):303–333.

Bryk, A. S., Sebring, P. B., Allenswoth, E., Luppescu, S., and Easton, J. Q. (2010). Organizing schools for improvement: Lessons from Chicago. *University of Chicago Press*.

Campbell, S. L. and Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*, 55(6):1233–1267.

Chaplin, D., Gill, B., Thompkins, A., and Miller, H. (2014). Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the pittsburgh public schools. rel 2014-024. *Regional Educational Laboratory Mid-Atlantic*.

Cherng, H.-Y. S., Halpin, P. F., and Rodriguez, L. A. (2022). Teaching bias? relations between teaching quality and classroom demographic composition. *American Journal of Education*, 128(2):171–201.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–79.

Chicago Public Schools (2014). Cps framework for teaching: Companion guide: Version 2.0 – august 2014.

Chicago Public Schools (2019). Reach students: Teacher performance evaluation.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of human Resources*, 41(4):778–820.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics*, 86(1):195–210.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources*, 42(3):528–554.

Delgado, W. (2025). Disparate teacher effects, comparative advantage, and match quality. *Economics of Education Review*, 106:102648.

District of Columbia Public Schools (2022). Impact annual reference guide 2022-2023.

Doherty, K. and Jacobs, S. (2013). Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice. national council on teacher quality.

Egalite, A. J., Kisida, B., and Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review*, 45:44 – 52.

Gershenson, S., Hart, C., Hyman, J., Lindsay, C. A., and Papageorge, N. W. (2022). The long-run impacts of same-race teachers. *American Economic Journal: Economic Policy*, 14(4):300–342.

Goldhaber, D., Lavery, L., and Theobald, R. (2015). Uneven playing field? assessing the teacher quality gap between advantaged and disadvantaged students. *Educational researcher*, 44(5):293–307.

Grissom, J. A. and Bartanen, B. (2022). Potential race and gender biases in high-stakes teacher observations. *Journal of Policy Analysis and Management*, 41(1):131–161.

Harris, D. N. and Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, 40:183–204.

Hart, H., Young, C., Chen, A., Zou, A., and Allensworth, E. M. (2020). Supporting school improvement: Early findings from a reexamination of the" 5essentials" survey. research report. *University of Chicago Consortium on School Research*.

Jacob, B. A. and Lefgren, L. (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of labor Economics*, 26(1):101–136.

Kalogrides, D. and Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. *Educational Researcher*, 42(6):304–316.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Kane, T. J. and Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. In *Research Paper. MET Project. Bill and Melinda Gates Foundation*.

Kane, T. J., Taylor, E. S., Tyler, J. H., and Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, 46(3):587–613.

Kraft, M. A. and Papay, J. P. (2014). Can professional environments in schools promote teacher development? explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis*, 36(4):476–500.

Lankford, H., Loeb, S., and Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, 24(1):37–62.

Loeb, S., Soland, J., and Fox, L. (2014). Is a good teacher a good teacher for all? comparing value-added of teachers with their english learners and non-english learners. *Educational Evaluation and Policy Analysis*, 36(4):457–475.

McCaffrey, D. F., Staiger, D. O., and Lockwood, J. (2013). A composite estimator of effective teaching. In *Research Paper. MET Project. Bill and Melinda Gates Foundation.*

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, 48(1):163–193.

Paufler, N. A. and Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal*, 51(2):328–362.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2):247–252.

Rockoff, J. E. and Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from new york city. *Labour Economics*, 18(5):687–696.

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

Rothstein, J. (2017). Measuring the impacts of teachers: comment. *American Economic Review*, 107(6):1656–1684.

Sanders, W. L. and Horn, S. P. (1998). Research findings from the tennessee value-added assessment system (tvaas) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3):247–256.

Sartain, L., Stoelinga, S. R., and Brown, E. R. (2011). *Rethinking Teacher Evaluation in Chicago: Lessons Learned from Classroom Observations, Principal-Teacher Conferences, and District Implementation. Research Report.* ERIC.

Sartain, L., Zou, A., Gutiérrez, V., Shyjka, A., Hinton, E., Brown, E. R., and Easton, J. Q. (2020). Teacher evaluation in cps: Perceptions of reach implementation, five years in. research brief. *University of Chicago Consortium on School Research.*

Sporte, S. E. and Jiang, J. Y. (2016). *Teacher Evaluation in Practice: Year 3 Teacher and Administrator Perceptions of REACH. Research Brief.* ERIC.

Steinberg, M. P. and Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-nclb era. *Education Finance and Policy*, 11(3):340–359.

Steinberg, M. P. and Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, 38(2):293–317.

Whitehurst, G., Chingos, M. M., and Lindquist, K. M. (2014). Evaluating teachers with classroom observations. *Brown Center on Education Policy: Brookings Institute*.

Tables

Table 1: Summary Statistics of Analytic and Survey Samples

	Classroom observation sample			Survey Sample		
	Mean (1)	Std. Dev. (2)	Obs. (3)	Mean (4)	Std. Dev. (5)	Obs. (6)
Panel A: Teacher characteristics						
Number of years per teacher	2.79	(1.54)	10,934	3.32	(1.54)	3,345
Female	0.83	(0.37)	30,373	0.78	(0.42)	10,959
White non-Hispanic	0.47	(0.5)	29,929	0.48	(0.5)	10,827
Black non-Hispanic	0.26	(0.44)	29,929	0.28	(0.45)	10,827
Hispanic	0.21	(0.41)	29,929	0.17	(0.38)	10,827
Age	40.11	(10.44)	30,337	40.21	(10.07)	10,947
Years of experience	9.26	(7.86)	29,816	9.14	(7.57)	10,787
Tenure status	0.74	(0.44)	30,479	0.77	(0.42)	10,991
Bachelor's degree	0.96	(0.19)	29,419	0.97	(0.17)	10,651
Master's degree	0.70	(0.46)	29,419	0.73	(0.44)	10,651
Doctorate degree	0.01	(0.08)	29,419	0.01	(0.09)	10,651
National Board Certification	0.08	(0.27)	29,817	0.09	(0.29)	10,788
Other teaching certification	0.98	(0.14)	29,817	0.99	(0.12)	10,788
Panel B: Classroom Characteristics						
Student-to-teacher ratio (class size)	62.10	(39.63)	30,479	85.94	(30.18)	10,991
Perc. male	0.50	(0.08)	30,479	0.50	(0.06)	10,991
Perc. White non-Hispanic	0.11	(0.19)	30,479	0.09	(0.17)	10,991
Perc. Black non-Hispanic	0.38	(0.42)	30,479	0.38	(0.42)	10,991
Perc. Hispanic	0.46	(0.39)	30,479	0.47	(0.39)	10,991
Perc. free lunch	0.85	(0.22)	30,479	0.87	(0.2)	10,991
Perc. special ed	0.11	(0.06)	30,479	0.12	(0.06)	10,991
Test scores (lagged)	0.03	(0.51)	30,479	0.04	(0.44)	10,991
Mean GPA (lagged)	3.10	(0.42)	30,479	3.02	(0.39)	10,991
Attendance rate (lagged)	0.95	(0.02)	30,478	0.96	(0.01)	10,991
Perc. suspended (lagged)	0.05	(0.08)	30,479	0.09	(0.09)	10,991
Perc. repeater	0.02	(0.04)	30,479	0.01	(0.03)	10,991
Classroom quality index	0.00	(0.82)	29,078	-0.07	(0.73)	10,990
Panel C: Teacher quality measures						
Classroom observation score	3.20	(0.48)	30,479	3.22	(0.48)	10,991
Survey score	3.18	(0.24)	10,991	3.18	(0.24)	10,991
Value added mean	0.06	(0.74)	20,435	0.06	(0.74)	8,674
Performance task summative	44.49	(22.43)	28,608	39.57	(16.34)	10,277

Notes: This table reports descriptive statistics for teacher demographics, classroom characteristics, and teacher quality measures from Chicago Public Schools between the 2011–12 and 2016–17 school years. The observation sample includes teachers with classroom observation ratings who are linked to at least five students with non-missing baseline test scores. The survey sample further restricts to teachers with matched student survey data. Classroom characteristics are classroom-level averages of individual student attributes (e.g., baseline test scores, attendance, demographics). Teacher quality measures include observation scores, survey ratings, and value-added estimates standardized by year. All continuous variables are reported in their natural units; standardized variables have mean zero and standard deviation one.

Table 2: Effects of Classroom Characteristics on Classroom Observation and Survey Scores

	Classroom observation score						Survey score					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Classroom quality (std)	0.172***	0.151***	0.094***	0.068***	0.077***		0.193***	0.180***	0.232***	0.152***	0.155***	
	(0.011)	(0.010)	(0.009)	(0.013)	(0.012)		(0.024)	(0.024)	(0.027)	(0.042)	(0.039)	
Perc. male	-0.357***	-0.275***	-0.227***	-0.123		-0.153**	0.166	0.219	0.212	-0.027		-0.140
	(0.078)	(0.074)	(0.065)	(0.078)		(0.076)	(0.163)	(0.161)	(0.157)	(0.219)		(0.218)
Perc. Black	-0.523***	-0.418***	-0.207**	-0.217		-0.239*	0.414***	0.407***	-0.512*	-0.921**		-1.073***
	(0.023)	(0.027)	(0.092)	(0.133)		(0.131)	(0.040)	(0.049)	(0.285)	(0.400)		(0.399)
Perc. non-Hispanic non-Black	0.225***	0.191***	0.049	-0.006		0.029	0.259***	0.308***	0.287	0.109		0.194
	(0.054)	(0.054)	(0.091)	(0.114)		(0.110)	(0.099)	(0.100)	(0.273)	(0.386)		(0.385)
Perc. free lunch	-0.235***	-0.191***	-0.295***	-0.151		-0.231**	0.680***	0.667***	0.578***	0.928***		0.800***
	(0.061)	(0.059)	(0.082)	(0.103)		(0.098)	(0.119)	(0.118)	(0.203)	(0.259)		(0.257)
Perc. special ed	-0.054	-0.084	-0.247***	-0.040		-0.175*	0.168	0.195	-0.282	-0.218		-0.506*
	(0.104)	(0.098)	(0.088)	(0.105)		(0.101)	(0.190)	(0.188)	(0.182)	(0.287)		(0.280)
Log class size	0.006	0.013	-0.008	-0.036**		-0.039**	-0.265***	-0.288***	-0.209***	-0.206***		-0.207***
	(0.012)	(0.011)	(0.010)	(0.018)		(0.018)	(0.031)	(0.031)	(0.032)	(0.061)		(0.061)
Teacher characteristics		Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes
School f.e.			Yes						Yes			
Teacher-school f.e.				Yes	Yes	Yes				Yes	Yes	Yes
R2	0.16	0.24	0.41	0.84	0.84	0.84	0.04	0.06	0.29	0.68	0.68	0.68
N (teachers)	10,354	10,354	10,354	10,354	10,354	10,934	4,066	4,066	4,066	4,066	4,066	4,067
N	29,078	29,078	29,078	29,078	29,078	30,479	10,990	10,990	10,990	10,990	10,990	10,991

Notes: Each column reports estimates from regressions of standardized teacher performance measures (observation or survey scores) on classroom characteristics. Standard errors are reported in parentheses and clustered at the teacher level unless teacher-by-school fixed effects are included, in which case robust standard errors are reported. Columns 1–6 use the observation sample; Columns 7–12 use the survey sample. All models include year fixed effects. Column 1 includes the classroom quality index and classroom demographics; Column 2 adds teacher characteristics (gender, race/ethnicity, age, experience, tenure, degree status, National Board Certification, and certification type); Column 3 adds school fixed effects; Column 4 includes teacher-by-school fixed effects; Columns 5 and 6 separately estimate specifications including only the classroom quality index or only classroom demographics. The same specification order applies to Columns 7–12 for the survey outcomes. Both the dependent variable and the classroom quality index are standardized by year (mean = 0, SD = 1). The classroom quality index is the classroom-level average of a student index constructed via principal component analysis of baseline test scores and behavioral variables. Asterisks denote significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table 3: Heterogeneity: Effects of Classroom Characteristics on Classroom Observation and Survey Scores across Teacher Subsamples

Sample	All	Female	Male	White	Black	Hispanic	Experienced teachers	Small class	Elementary	Middle
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Classroom observation score										
Classroom quality (std)	0.068***	0.065***	0.097**	0.065***	0.078***	0.038	0.069***	0.066***	0.054***	0.106***
	(0.013)	(0.013)	(0.041)	(0.020)	(0.023)	(0.028)	(0.017)	(0.019)	(0.015)	(0.027)
R2	0.84	0.84	0.84	0.84	0.81	0.81	0.82	0.87	0.86	0.85
N (teachers)	10,354	8,623	1,731	4,758	2,636	2,068	6,043	5,546	7,042	4,653
N	29,078	24,050	5,028	13,270	7,522	5,916	17,664	10,995	16,341	12,737
Panel B: Survey score										
Classroom quality (std)	0.152***	0.148***	0.173**	0.223***	0.059	0.179*	0.145***			
	(0.042)	(0.048)	(0.085)	(0.067)	(0.067)	(0.098)	(0.050)			
R2	0.68	0.68	0.68	0.69	0.69	0.64	0.67			
N (teachers)	4,066	3,188	878	1,894	1,161	661	2,436			
N	10,990	8,567	2,423	5,203	3,038	1,856	6,865			
Class demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher-school f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each column reports estimates from regressions of standardized teacher performance measures (observation or survey scores) on classroom characteristics across different teacher subsamples. Robust standard errors are reported in parentheses. All specifications control for classroom demographics, teacher characteristics, year fixed effects, and teacher-by-school fixed effects. Panel A reports estimates for classroom observation scores; Panel B reports estimates for student survey scores. Column 1 includes the full sample. Columns 2–3 split the sample by teacher gender; Columns 4–6 by race/ethnicity (White, Black, Hispanic); Column 7 restricts to teachers with five or more years of experience; Column 8 restricts to teachers with class sizes of 35 students or fewer; and Columns 9–10 separate elementary and middle school teachers based on the classroom's modal grade. All dependent variables and the classroom quality index are standardized by year (mean = 0, SD = 1). The classroom quality index is the classroom-level average of a student index constructed using principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table 4: Heterogeneity: Effects of Classroom Characteristics on Classroom Observation and Survey Scores by Baseline Teacher Quality

	Lagged observation sample			Lagged VA sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Classroom observation scores						
Classroom quality (std)	0.053***	0.053***	0.186	0.064***	0.058***	0.275*
	(0.017)	(0.017)	(0.122)	(0.021)	(0.021)	(0.140)
L.Observation score (std)		0.033**	0.033**			
		(0.015)	(0.014)			
Classroom quality # L.Observation score (std)			0.032*			
			(0.018)			
L.Value added mean (std)					0.027***	0.027***
					(0.009)	(0.009)
Classroom quality # L.Value added mean (std)						0.002
						(0.010)
R2	0.88	0.88	0.88	0.89	0.89	0.89
N (teachers)	6,817	6,817	6,817	5,342	5,342	5,342
N	17,264	17,264	17,264	12,854	12,854	12,854
Panel B: Survey scores						
Classroom quality (std)	0.093*	0.092	0.098	0.133**	0.127**	-0.010
	(0.056)	(0.056)	(0.312)	(0.061)	(0.063)	(0.339)
L.Observation score (std)		0.023	0.024			
		(0.031)	(0.031)			
Classroom quality # L.Observation score (std)			0.007			
			(0.039)			
L.Value added mean (std)					0.016	0.013
					(0.022)	(0.022)
Classroom quality # L.Value added mean (std)						-0.020
						(0.032)
R2	0.70	0.70	0.70	0.72	0.72	0.72
N (teachers)	2,992	2,992	2,992	2,486	2,486	2,486
N	7,282	7,282	7,282	5,712	5,712	5,712
Class demographics	Yes	Yes	Yes	Yes	Yes	Yes
Teacher characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Teacher-school f.e.	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each column reports estimates from regressions of standardized teacher performance measures (observation or survey scores) on the classroom quality index, a baseline teacher quality measure, and their interaction. Robust standard errors are reported in parentheses. All specifications include classroom demographics, teacher characteristics, year fixed effects, and teacher-by-school fixed effects. Panel A reports results for classroom observation scores; Panel B reports results for student survey scores. Columns 1–3 restrict the sample to teachers with prior-year observation scores as the baseline teacher quality measure. Columns 4–6 restrict to teachers with prior-year value-added scores. Each column additionally controls for baseline teacher quality and the interaction between baseline teacher quality and classroom quality. All dependent variables, the classroom quality index, and baseline teacher quality measures are standardized by year (mean = 0, SD = 1). The classroom quality index is defined as the classroom-level average of a student index derived via principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table 5: Exogeneity Checks: Placebo Tests for the Effects of Classroom Characteristics on Prior and Future-Year Observation and Survey Scores

	t-3	*t-2*	*t-1*	*t*	*t+1*	*t+2*	*t+3*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Classroom observation score							
Classroom quality (std)	-0.002	0.004	-0.000	0.068***	0.033**	-0.019	0.026
	(0.072)	(0.026)	(0.020)	(0.013)	(0.016)	(0.021)	(0.035)
R2	0.87	0.84	0.82	0.84	0.87	0.90	0.95
N (teachers)	4,569	5,403	6,817	10,354	6,838	5,425	4,571
N	7,064	11,495	17,264	29,078	17,270	11,501	7,069
Panel B: Survey score							
Classroom quality (std)	-0.011	0.091	0.079	0.152***	0.021	-0.062	-0.100
	(0.127)	(0.076)	(0.053)	(0.042)	(0.049)	(0.072)	(0.144)
R2	0.82	0.73	0.68	0.68	0.71	0.76	0.83
N (teachers)	1,977	2,438	3,035	4,066	2,991	2,447	1,955
N	3,023	4,936	7,313	10,990	7,276	4,947	2,995
Class demographics	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher-school f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Each column reports estimates from regressions of prior, current, and future-year teacher performance ratings (observation and survey scores) on classroom characteristics. Robust standard errors are reported in parentheses. All specifications include classroom demographics, teacher characteristics, year fixed effects, and teacher-by-school fixed effects. Panel A reports results for classroom observation scores; Panel B reports results for student survey scores. The dependent variable corresponds to the teacher's observation or survey score in year $t-3$ to $t+3$. The dependent variables and the classroom quality index are standardized by year (mean = 0, SD = 1). The classroom quality index is the classroom-level average of a student index derived via principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table 6: Effects of Classroom Characteristics on Teacher Productivity Measures

	Performance task sample			VA sample		
	Observation score	Performance task summative		Observation score	VA mean	
	(1)	(2)	(3)	(4)	(5)	(6)
Classroom quality (std)	0.068***	-2.066***	-1.037**	0.083***	0.131***	-0.148***
	(0.013)	(0.246)	(0.495)	(0.018)	(0.011)	(0.023)
Class demographics	Yes	Yes	Yes	Yes	Yes	Yes
Teacher characteristics	Yes		Yes	Yes		Yes
Teacher-school f.e.	Yes		Yes	Yes		Yes
R2	0.84	0.08	0.72	0.84	0.02	0.61
N (teachers)	10,017	10,017	10,017	7,373	7,373	7,373
N	27,236	27,236	27,236	20,435	20,435	20,435

Notes: Each column reports estimates from regressions of classroom observation, performance task, or value-added scores on classroom characteristics. Standard errors are reported in parentheses and clustered at the teacher level in specifications without teacher-by-school fixed effects; robust standard errors are reported when teacher-by-school fixed effects are included. Columns 1–3 include teachers with performance task data; Columns 4–6 include teachers with value-added data. All models include year fixed effects and classroom demographics, and some specifications additionally include teacher characteristics and teacher-by-school fixed effects. Both the dependent variables and the classroom quality index are standardized by year (mean = 0, SD = 1). The classroom quality index is the classroom-level average of a student index derived via principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table 7: Teacher Characteristics and the Effects of Adjusting Observation Scores for Classroom Composition

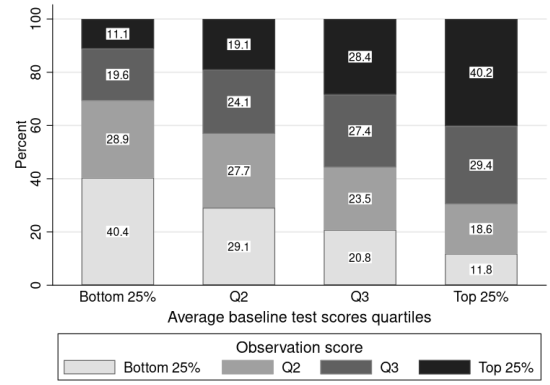
	Two-group tournament			Three-group tournament			Quality-adjusted score		
	Class obs ranking change	Exit bottom 5%	Exit top 5%	Class obs ranking change	Exit bottom 5%	Exit top 5%	Class obs ranking change	Exit bottom 5%	Exit top 5%
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Male	0.50**	-0.06**	-0.02	0.58**	-0.08***	-0.06*	0.25	-0.05**	-0.10**
	(0.24)	(0.03)	(0.04)	(0.26)	(0.03)	(0.04)	(0.24)	(0.02)	(0.05)
Black	7.16***	0.07***	-0.07	8.87***	0.08***	-0.09**	8.29***	0.11***	-0.16***
	(0.21)	(0.03)	(0.05)	(0.24)	(0.03)	(0.05)	(0.24)	(0.03)	(0.06)
Hispanic	2.13***	0.01	-0.04	2.27***	-0.04	-0.05	2.15***	-0.03	-0.11***
	(0.23)	(0.04)	(0.03)	(0.25)	(0.03)	(0.03)	(0.22)	(0.03)	(0.04)
Other race (non-White)	-0.45	0.04	0.10	-0.26	-0.02	0.11	-0.53	0.02	0.19**
	(0.39)	(0.05)	(0.07)	(0.44)	(0.05)	(0.07)	(0.45)	(0.05)	(0.09)
Age	0.02	-0.00***	0.00	0.03**	-0.00*	0.00*	0.04***	-0.00	0.00
	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Years of experience	-0.07***	-0.00	-0.00	-0.09***	-0.00**	-0.00	-0.09***	-0.00*	0.00
	(0.02)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)
Tenured	-0.06	0.02	-0.08	0.06	0.04	-0.07	0.14	0.01	-0.06
	(0.22)	(0.03)	(0.05)	(0.24)	(0.03)	(0.05)	(0.24)	(0.03)	(0.06)
Graduate degree	0.18	0.05*	-0.08**	0.16	0.06**	-0.07**	0.27	0.04*	-0.04
	(0.20)	(0.03)	(0.04)	(0.22)	(0.03)	(0.04)	(0.21)	(0.03)	(0.04)
National Board Certification	-0.70**	-0.03	0.06	-0.51	-0.12	0.05	-0.29	-0.05	0.09*
	(0.33)	(0.08)	(0.04)	(0.36)	(0.08)	(0.04)	(0.36)	(0.06)	(0.05)
No teaching certification	-0.12	-0.07	0.06	-0.17	-0.03	-0.09	-0.18	-0.02	-0.12
	(0.39)	(0.05)	(0.10)	(0.44)	(0.05)	(0.10)	(0.47)	(0.05)	(0.12)
R2	0.09	0.03	0.03	0.11	0.05	0.03	0.11	0.04	0.06
N	30,479	1,558	1,432	30,479	1,558	1,432	29,078	1,495	1,374

Notes: Each column reports regressions of measures of adjusted teacher rankings on teacher characteristics. Standard errors are reported in parentheses and clustered at the teacher level. The sample includes teachers in the classroom observation sample. The outcome variables are defined as follows: Ranking change is the difference between a teacher's percentile ranking in the adjusted and unadjusted observation score distributions (higher values indicate improvement under the adjustment). Exit bottom equals 1 if a teacher moves above the 5th percentile in the adjusted distribution, conditional on being below it in the unadjusted distribution. Exit top equals 1 if a teacher moves below the 95th percentile in the adjusted distribution, conditional on being above it in the unadjusted distribution. Columns 1–3 adjust rankings by dividing teachers into two groups based on whether their classroom quality index is below or above the median; Columns 4–6 divide teachers into three groups (low, medium, high); and Columns 7–9 use a regression-based adjustment for classroom characteristics. The omitted racial group is White, non-Hispanic teachers. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

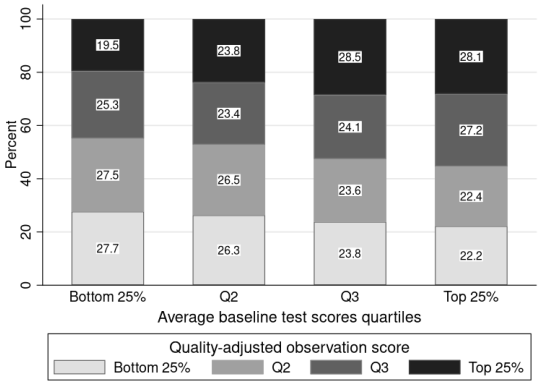
Figures

Figure 1: Conditional Distribution of Unadjusted and Quality-Adjusted Classroom Observation and Survey Scores given Average Baseline Test Scores

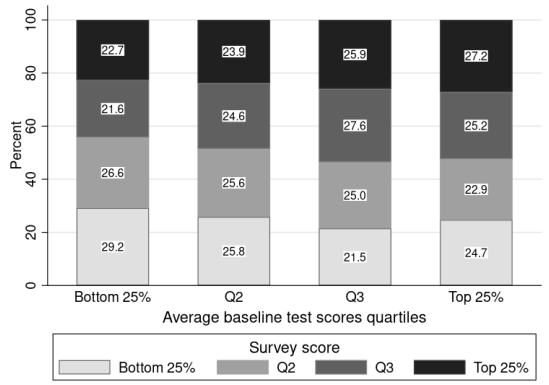
(a) Unadjusted observation score



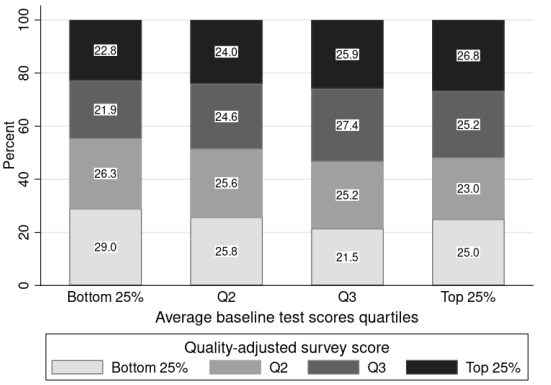
(b) Quality-adjusted observation score



(c) Unadjusted survey score



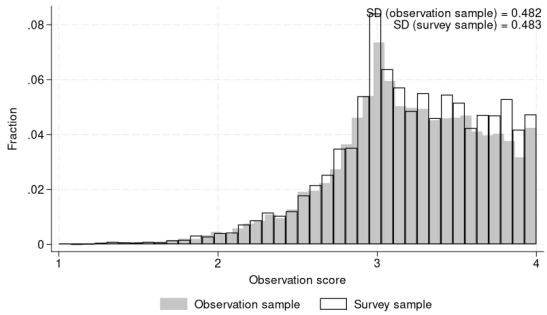
(d) Quality-adjusted survey score



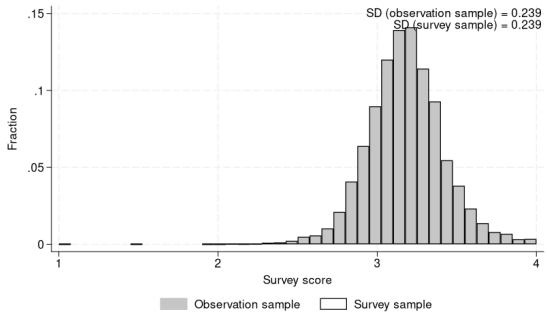
Notes: Figures show the conditional distribution of classroom observation and survey scores by quartiles of average classroom baseline test scores. Each figure divides classrooms into four equal-sized groups based on their average baseline test scores and plots the proportion of teachers in each performance-score quartile. Panels A and C display unadjusted observation and survey scores; Panels B and D display scores adjusted for classroom quality and year fixed effects. Adjusted scores are residuals from regressions of raw scores on the classroom quality index and year fixed effects.

Figure 2: Distribution of Teacher Performance Ratings and Selected Classroom Characteristics (Between-Teacher Variation)

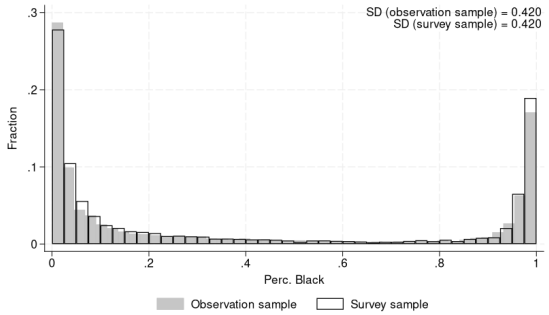
(a) Classroom observation score



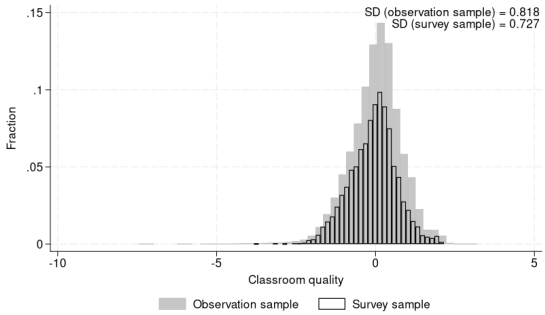
(b) Survey score



(c) Proportion of Black students



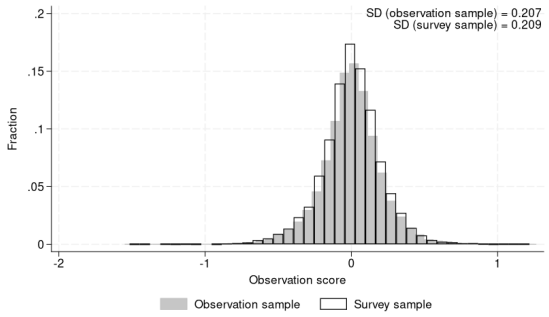
(d) Classroom quality index



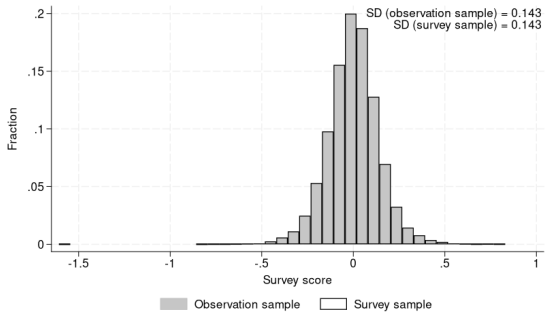
Notes: Figures show between-teacher distributions of classroom observation scores, student survey scores, and selected classroom characteristics. Each panel displays the overall variation across teachers. The classroom quality index is the classroom-level average of a student index derived from principal component analysis of baseline test scores and behavioral variables.

Figure 3: Distribution of Teacher Performance Ratings and Selected Classroom Characteristics (Within-Teacher Variation)

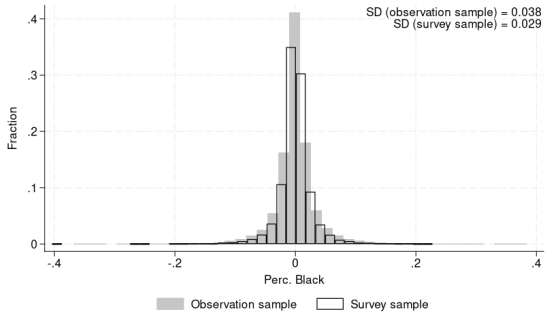
(a) Classroom observation score



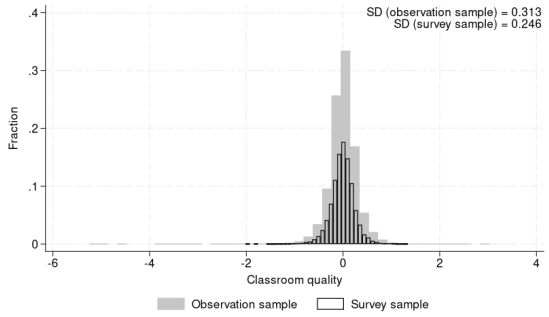
(b) Survey score



(c) Proportion of Black students



(d) Classroom quality index



Notes: Figures show within-teacher distributions of classroom observation scores, student survey scores, and selected classroom characteristics. Each variable is residualized after partialling out teacher-by-school and year fixed effects, capturing within-teacher variation over time. The classroom quality index is the classroom-level average of a student index derived from principal component analysis of baseline test scores and behavioral variables.

Online Appendix Materials

A Appendix Tables and Figures

Appendix Tables

Table A.1: Classroom Observation Domains and Components

Domain	Component	Rubric for distinguished rating (4 out of 4)
1. Planning and preparation	1a. Demonstrating knowledge of content and pedagogy	Teacher demonstrates knowledge of the relevant content standards within the grade level and across grade levels, as well as how these standards relate to other disciplines.
	1b. Demonstrating knowledge of students	The teacher demonstrates an understanding of the active nature of student learning and attains information about levels of development for individual students.
	1c. Selecting learning objectives	Learning objectives are standards-based, clear, written in the form of student learning outcomes, aligned to methods of assessment, and varied in whatever way is needed to account for individual students' needs.
	1d. Designing coherent instruction	Teacher coordinates in-depth knowledge of content, students, and resources (including technology) to design units and lessons.
	1e. Designing student assessment	The plan for student assessment is aligned with the standards-based learning objectives identified for the unit and lesson.
2. Classroom environment	2a. Creating an environment of respect and rapport	Patterns of classroom interactions, both between the teacher and students and among students, are highly respectful, reflecting genuine warmth and caring.
	2b. Establishing a culture for learning	The teacher creates a classroom culture that reflects a shared belief in the importance of learning and hard work.
	2c. Managing classroom procedures	Effective classroom routines and procedures maximize instructional time. The teacher orchestrates the environment so that students contribute to the management of instructional groupings, transitions, and/or the handling of materials and supplies without disruption of learning.
	2d. Managing student behavior	Teacher and students establish and implement standards of conduct. Students follow the standards of conduct and self-monitor their behaviors.
3. Instruction	3a. Communicating with students	Teacher clearly communicates standards-based learning objective(s). Teacher guides students to articulate the relevance of the objective(s) to learning.
	3b. Using questioning and discussion techniques	Teacher uses a variety of low- and high-level, open-ended, and developmentally appropriate questions to challenge students cognitively, advance high level thinking and discourse, and promote metacognition.
	3c. Engaging students in learning	Tasks align with standards-based learning objectives and are tailored so virtually all students are intellectually engaged in challenging content. Tasks and text are complex and promote student engagement through inquiry and choice.
	3d. Using assessment in instruction	Teacher fully integrates formative assessment into instruction, and uses it to monitor progress, and to check for understanding for individual students.
	3e. Demonstrating flexibility and responsiveness	Teacher seizes opportunities to enhance learning, building on a spontaneous world or local event and/or student interests.
4. Professional responsibilities	4a. Reflecting on teaching and learning	Teacher makes an accurate assessment of a lesson's or unit's effectiveness and the extent to which it achieved its objective and its impact on student learning, citing many specific examples and evidence.
	4b. Maintaining accurate records	Teacher has a detailed system for maintaining information on student completion of assignments, student progress in learning, and non- instructional records, requiring no monitoring for errors.
	4c. Communicating with families	Teacher frequently communicates with families to convey information about class and individual activities, individual student's progress and to solicit and utilize the family's support in student learning.
	4d. Growing and developing professionally	Teacher initiates opportunities for professional growth and makes a systematic effort to enhance content knowledge and pedagogical skill of self and colleagues.
	4e. Demonstrating professionalism	Teacher has the highest standards of integrity, always holds student and required school information confidential, and is honest in professional and student/family interactions.

Source: Chicago Public Schools (2014).
Notes: This table lists the rubric domains and component indicators used to construct classroom observation scores. The final column provides the opening sentences of the "Distinguished" (highest) rating descriptor for each domain.

Table A.2: Example of a Rubric Associated with the Instruction Domain of REACH

Component	*Unsatisfactory*	*Basic*	*Proficient*	*Distinguished*
2a: Creating an Environment of Respect and Rapport • *Teacher Interactions with Students* • *Student Interactions with Other Students*	Patterns of classroom interactions, both between the teacher and students and among students, are mostly negative and disrespectful. Interactions are insensitive and/or inappropriate to the ages and development of the students, and the context of the class. The net result of interactions has a negative impact on students emotionally and/or academically.	Patterns of classroom interactions, both between the teacher and students and among students, are generally respectful but may reflect occasional inconsistencies or incidences of disrespect. Some interactions are sensitive and/or appropriate to the ages and development of the students, and the context of the class. The net result of the interactions has a neutral impact on students emotionally and/or academically.	Patterns of classroom interactions, both between the teacher and students and among students, are friendly and demonstrate caring and respect. Interactions among students are generally polite and respectful. Interactions are sensitive and appropriate to the ages and development of the students, and to the context of the class. The net result of the interactions has a positive impact on students emotionally and academically.	Patterns of classroom interactions, both between the teacher and students and among students, are highly respectful, reflecting genuine warmth and caring. Students contribute to high levels of civility among all members of the class. Interactions are sensitive to students as individuals, appropriate to the ages and development of individual students, and to the context of the class. The net result of interactions is that of academic and personal connections among students and adults.

Source: Chicago Public Schools (2014). Notes: This table reproduces an excerpt from the REACH performance evaluation rubric used by Chicago Public Schools. It illustrates the scoring guide for Component 2a, *Creating an Environment of Respect and Rapport*, which belongs to Domain 2, *Classroom Environment*. The rubric describes observable behaviors corresponding to each performance level, ranging from Unsatisfactory to Distinguished.

Table A.3: Student Survey Questions Related to the Teacher and Classroom

Index	**Survey questions**
Peer support	A. How many students in your class. . . 1 Feel it is important to come to school every day? 2 Feel it is important to pay attention in class? 3 Think doing homework is important? 4 Try hard to get good grades?
Classroom rigor	B. How much do you disagree or agree with the following statements about your teacher in your class? My teacher. . . 1 Often connects what I am learning to life outside of the classroom. 2 Encourages students to share their ideas about things we are studying in class. 3 Often requires me to explain my answers. 4 Encourages us to consider different solutions or points of view. 5 Doesn't let students give up when the work gets hard.
	C. How often does the following occur? In my [TARGET] class, we talk about different solutions or points of view.
Academic press	D. How much do you disagree or agree with the following statements about your class? 1 This class really makes me think. 2 I'm really learning a lot in this class.
	E. To what extent do you disagree or agree with the following statements? In my class, my teacher. . . 1 Expects everyone to work hard. 2 Expects me to do my best all the time. 3 Wants us to become better thinkers, not just memorize things.
	F. In your class, how often. . . 1 Are you challenged? 2 Do you have to work hard to do well? 3 Does the teacher ask difficult questions on tests? 4 Does the teacher ask difficult questions in class?
Course clarity	G. How much do you disagree or agree with the following statements about your class? 1 I learn a lot from feedback on my work. 2 It's clear to me what I need to do to get a good grade. 3 The work we do in class is good preparation for the test. 4 The homework assignments help me to learn the course material. 5 I know what my teacher wants me to learn in this class.
Academic engagement	H. How much do you disagree or agree with the following statements about your class? 1 I usually look forward to this class. 2 I work hard to do my best in this class. 3 Sometimes I get so interested in my work I don't want to stop. 4 The topics we are studying are interesting and challenging.
Academic personalisms	I. How much do you disagree or agree with the following statements about your class? The teacher for this class. . . 1 Helps me catch up if I am behind. 2 Is willing to give extra help on schoolwork if I need it. 3 Notices if I have trouble learning something. 4 Gives me specific suggestions about how I can improve my work in this class. 5 Explains things in a different way if I don't understand something in class.
Classroom disruptions	J. How much do you disagree or agree with the following statement about your class? 1 I get distracted from my work by other students acting out in this class. (Reverse coded) 2 This class is out of control. (Reverse coded) 3 My classmates do not behave the way my teacher wants them to. (Reverse coded)

Notes: This table reports the 5Essentials student survey items that are subject- and classroom-specific and consistently asked across years. Each student was randomly asked about math, English, or science. Items are grouped into seven indices listed in the first column. Response options (numerical values in parentheses) are: A: (1) None, (2) A few, (3) About half, (4) Most, (5) All. B, D, E, G, H, I, J: (1) Strongly disagree, (2) Disagree, (3) Agree, (4) Strongly agree. C: (1) Very little, (2) Some, (3) Quite a bit, (4) A great deal. F: (1) Never, (2) Once in a while, (3) Most of the time, (4) All the time. Items in J (classroom disruptions) are reverse coded so that higher values indicate more positive outcomes.

Table A.4: Weights and Explained Variation of the Components Resulting from Principal Component Analysis of Student-level Achievement and Behaviors

	Component 1	Component 2	Component 3	Component 4	Component 5
Test score	0.55	-0.09	0.51	-0.14	-0.64
GPA	0.60	-0.01	0.25	-0.14	0.74
Attendance rate	0.40	0.21	-0.25	0.85	-0.08
Suspended	-0.34	-0.49	0.62	0.48	0.16
Repeater	-0.24	0.84	0.48	0.05	0.05
Explained variance	39%	19%	18%	16%	8%

Notes: Table reports weights of the components resulting from conducting principal component analysis on student-level baseline test scores, GPA, attendance rate, suspended indicator, and repeater indicator. Proportion of explained variance of each component is shown in the last row.

Table A.5: Correlation between Classroom Observation and Survey Scores with Various Teacher Quality Measures

	Classroom observation score	Survey score	Quality-adjusted classroom observation score	Quality-adjusted survey score
	(1)	(2)	(3)	(4)
Panel A: Class observation domains				
Classroom observation score	1.00	0.24	0.91	0.19
Planning and preparation	0.92	0.20	0.84	0.15
Classroom environment	0.90	0.26	0.83	0.21
Instruction	0.93	0.23	0.85	0.18
Professional responsibilities	0.86	0.18	0.79	0.14
Panel B: Survey indexes				
Survey score	0.24	1.00	0.20	0.96
Peer support	0.13	0.76	0.10	0.73
Classroom rigor	0.23	0.88	0.20	0.88
Academic press	0.17	0.85	0.16	0.84
Course clarity	0.15	0.87	0.15	0.88
Academic engagement	0.08	0.84	0.12	0.83
Academic personalism	0.12	0.85	0.13	0.85
Classroom disruptions	0.36	0.71	0.24	0.60
Panel C: Student growth measures				
Performance task summative	0.07	-0.01	0.01	-0.06
Value added mean	0.22	0.17	0.20	0.18
Value added math	0.24	0.21	0.23	0.22
Value added reading	0.20	0.13	0.19	0.14

Notes: This table reports correlations between classroom observation scores, student survey scores, and external measures of teacher quality. Columns 3 and 4 use observation and survey scores adjusted for classroom composition (residuals from regressions on the classroom quality index and year fixed effects).

Table A.6: Effects of Classroom Quality Components and Demographic Classroom Characteristics on Classroom Observation and Survey Scores

	Classroom observation score				Survey score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Classroom quality (1st component)	0.194***	0.077***	0.087***		0.173***	0.140***	0.141***	
	(0.011)	(0.014)	(0.013)		(0.024)	(0.042)	(0.039)	
Classroom quality (2nd component)	0.060***	0.038**	0.044**		0.254***	0.204***	0.219***	
	(0.016)	(0.019)	(0.020)		(0.051)	(0.062)	(0.062)	
Classroom quality (3rd component)	0.097***	0.016	0.017		0.101**	0.046	0.041	
	(0.020)	(0.024)	(0.023)		(0.049)	(0.070)	(0.068)	
Perc. male	-0.332***	-0.119		-0.164**	0.168	-0.002		-0.140
	(0.078)	(0.078)		(0.078)	(0.163)	(0.219)		(0.218)
Perc. Black	-0.518***	-0.212		-0.269**	0.430***	-0.895**		-1.073***
	(0.024)	(0.133)		(0.133)	(0.043)	(0.400)		(0.399)
Perc. non-Hispanic non-Black	0.195***	-0.013		0.033	0.252**	0.107		0.194
	(0.054)	(0.115)		(0.114)	(0.100)	(0.386)		(0.385)
Perc. free lunch	-0.137**	-0.139		-0.217**	0.676***	0.903***		0.800***
	(0.062)	(0.104)		(0.103)	(0.122)	(0.261)		(0.257)
Perc. special ed	0.067	-0.014		-0.158	0.222	-0.220		-0.506*
	(0.105)	(0.107)		(0.104)	(0.193)	(0.292)		(0.280)
Log class size	0.012	-0.033*		-0.038**	-0.255***	-0.202***		-0.207***
	(0.012)	(0.018)		(0.018)	(0.031)	(0.060)		(0.061)
Teacher characteristics		Yes	Yes	Yes		Yes	Yes	Yes
Teacher-school f.e.		Yes	Yes	Yes		Yes	Yes	Yes
R2	0.17	0.84	0.84	0.84	0.04	0.68	0.68	0.68
N (teachers)	10,354	10,354	10,354	10,354	4,066	4,066	4,066	4,066
N	29,078	29,078	29,078	29,078	10,990	10,990	10,990	10,990

Notes: Each column reports estimates from regressions of standardized teacher performance measures (observation or survey scores) on classroom quality indexes and demographic characteristics. Standard errors are reported in parentheses and clustered at the teacher level in specifications without teacher-by-school fixed effects; robust standard errors are reported when teacher-by-school fixed effects are included. Columns 1–4 use the observation sample; Columns 5–8 use the survey sample. All models include year fixed effects. Column 1 includes the classroom quality indexes, which are the first three components of the principal component analysis of student baseline achievement and behaviors, and classroom demographics; Column 2 adds teacher characteristics (gender, race/ethnicity, age, experience, tenure, degree status, National Board Certification, and certification type) and teacher-by-school fixed effects; Columns 3–4 report analogous specifications. Columns 5–6 estimate models including only the classroom quality indexes or only demographics; Columns 7–8 report analogous specifications for the survey sample. The classroom quality indexes and the dependent variables are standardized by year (mean = 0, SD = 1). Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table A.7: Effects of Achievement, Behavioral, and Demographic Classroom Characteristics on Classroom Observation and Survey Scores

	Classroom observation score				Survey score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test scores (lagged) (std)	0.162***	0.034**	0.040***		0.152***	0.109**	0.100**	
	(0.011)	(0.014)	(0.013)		(0.024)	(0.049)	(0.045)	
Mean GPA (lagged) (std)	0.048***	0.027**	0.032***		0.022	0.054	0.056	
	(0.009)	(0.011)	(0.011)		(0.020)	(0.035)	(0.035)	
Attendance rate (lagged) (std)	0.030***	0.021**	0.020**		0.043**	-0.009	-0.009	
	(0.008)	(0.009)	(0.009)		(0.020)	(0.029)	(0.029)	
Perc. suspended (lagged)	-0.506***	-0.167	-0.204		-0.781***	-0.909***	-0.975***	
	(0.107)	(0.127)	(0.127)		(0.165)	(0.265)	(0.263)	
Perc. repeater	0.096	0.146	0.155		1.641***	0.899*	0.865*	
	(0.138)	(0.136)	(0.136)		(0.384)	(0.481)	(0.470)	
Perc. male	-0.300***	-0.115		-0.153**	0.195	0.026		-0.140
	(0.075)	(0.076)		(0.076)	(0.163)	(0.219)		(0.218)
Perc. Black	-0.525***	-0.182		-0.239*	0.399***	-0.890**		-1.073***
	(0.024)	(0.131)		(0.131)	(0.043)	(0.399)		(0.399)
Perc. non-Hispanic non-Black	0.178***	-0.015		0.029	0.234**	0.028		0.194
	(0.052)	(0.110)		(0.110)	(0.102)	(0.391)		(0.385)
Perc. free lunch	-0.048	-0.175*		-0.231**	0.730***	0.879***		0.800***
	(0.059)	(0.099)		(0.098)	(0.123)	(0.260)		(0.257)
Perc. special ed	0.251**	-0.041		-0.175*	0.369*	-0.115		-0.506*
	(0.102)	(0.105)		(0.101)	(0.194)	(0.300)		(0.280)
Log class size	-0.000	-0.037**		-0.039**	-0.265***	-0.206***		-0.207***
	(0.012)	(0.018)		(0.018)	(0.031)	(0.060)		(0.061)
Teacher characteristics		Yes	Yes	Yes		Yes	Yes	Yes
Teacher-school f.e.		Yes	Yes	Yes		Yes	Yes	Yes
R2	0.17	0.84	0.84	0.84	0.04	0.68	0.68	0.68
N (teachers)	10,933	10,933	10,933	10,934	4,067	4,067	4,067	4,067
N	30,478	30,478	30,478	30,479	10,991	10,991	10,991	10,991

Notes: Each column reports estimates from regressions of standardized teacher performance measures (observation or survey scores) on classroom achievement, behavioral, and demographic characteristics. Standard errors are reported in parentheses and clustered at the teacher level in specifications without teacher-by-school fixed effects; robust standard errors are reported when teacher-by-school fixed effects are included. Columns 1–4 use the observation sample; Columns 5–8 use the survey sample. All models include year fixed effects. Column 1 includes the full vector of classroom characteristics; Column 2 adds teacher characteristics (gender, race/ethnicity, age, experience, tenure, degree status, National Board Certification, and certification type) and teacher-by-school fixed effects; Columns 3–4 report analogous specifications. Columns 5–6 estimate models including only achievement/behavioral characteristics or only demographics; Columns 7–8 report analogous specifications for the survey sample. Baseline test scores, GPA, attendance rate, and the dependent variables are standardized by year (mean = 0, SD = 1). Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table A.8: Effects of Classroom Characteristics on Classroom Observation's Domains

	Planning and preparation		Classroom environment		Instruction		Professional responsibilities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Classroom quality (std)	0.158***	0.049***	0.173***	0.085***	0.166***	0.065***	0.125***	0.054***
	(0.011)	(0.014)	(0.011)	(0.014)	(0.010)	(0.014)	(0.010)	(0.016)
Perc. male	-0.287***	-0.070	-0.422***	-0.114	-0.321***	-0.062	-0.286***	-0.246**
	(0.079)	(0.086)	(0.079)	(0.081)	(0.079)	(0.081)	(0.080)	(0.102)
Perc. Black	-0.447***	-0.034	-0.528***	-0.332**	-0.445***	-0.234*	-0.475***	-0.094
	(0.023)	(0.146)	(0.023)	(0.139)	(0.023)	(0.140)	(0.022)	(0.164)
Perc. non-Hispanic non-Black	0.201***	0.086	0.136***	-0.044	0.216***	0.009	0.227***	-0.047
	(0.056)	(0.127)	(0.053)	(0.124)	(0.056)	(0.124)	(0.049)	(0.154)
Perc. free lunch	-0.247***	-0.156	-0.110*	-0.120	-0.283***	-0.132	-0.216***	-0.144
	(0.063)	(0.114)	(0.060)	(0.108)	(0.062)	(0.107)	(0.055)	(0.134)
Perc. special ed	-0.087	-0.027	-0.088	0.016	0.004	0.037	-0.078	-0.141
	(0.107)	(0.118)	(0.105)	(0.111)	(0.105)	(0.109)	(0.102)	(0.137)
Log class size	0.038***	-0.019	-0.024**	-0.045**	0.021*	-0.011	-0.011	-0.046**
	(0.012)	(0.020)	(0.012)	(0.019)	(0.012)	(0.019)	(0.011)	(0.022)
Teacher characteristics		Yes		Yes		Yes		Yes
Teacher-school f.e.		Yes		Yes		Yes		Yes
R2	0.13	0.80	0.14	0.82	0.14	0.82	0.12	0.71
N (teachers)	10,217	10,217	10,354	10,354	10,354	10,354	10,231	10,231
N	28,349	28,349	29,073	29,073	29,068	29,068	28,411	28,411

Notes: Each column reports estimates from regressions of standardized observation domain scores on classroom characteristics. Standard errors are reported in parentheses and clustered at the teacher level in specifications without teacher-by-school fixed effects; robust standard errors are reported when teacher-by-school fixed effects are included. The sample includes teachers in the observation sample. All models include year fixed effects; even-numbered columns additionally include teacher characteristics and teacher-by-school fixed effects. Both the dependent variables and the classroom quality index are standardized by year (mean = 0, SD = 1). The classroom quality index is the classroom-level average of a student index derived via principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table A.9: Effects of Classroom Characteristics on Classroom Observation in the Survey Sample

	Classroom observation score					
	(1)	(2)	(3)	(4)	(5)	(6)
Classroom quality (std)	0.271***	0.237***	0.155***	0.110***	0.117***	
	(0.021)	(0.020)	(0.021)	(0.028)	(0.027)	
Perc. male	-0.363**	-0.210	-0.176	-0.052		-0.133
	(0.156)	(0.147)	(0.136)	(0.167)		(0.167)
Perc. Black	-0.476***	-0.418***	-0.183	-0.262		-0.372
	(0.039)	(0.046)	(0.214)	(0.304)		(0.307)
Perc. non-Hispanic non-Black	0.254***	0.181**	-0.189	0.010		0.071
	(0.085)	(0.084)	(0.194)	(0.252)		(0.251)
Perc. free lunch	-0.070	-0.098	-0.525***	-0.244		-0.336*
	(0.104)	(0.099)	(0.155)	(0.201)		(0.201)
Perc. special ed	-0.088	-0.091	-0.382**	0.109		-0.098
	(0.185)	(0.172)	(0.164)	(0.201)		(0.195)
Log class size	-0.026	-0.060**	-0.088***	-0.072*		-0.073*
	(0.029)	(0.027)	(0.027)	(0.041)		(0.041)
Teacher characteristics		Yes	Yes	Yes	Yes	Yes
School f.e.			Yes			
Teacher-school f.e.				Yes	Yes	Yes
R2	0.19	0.28	0.47	0.85	0.85	0.85
N (teachers)	4,066	4,066	4,066	4,066	4,066	4,067
N	10,990	10,990	10,990	10,990	10,990	10,991

Notes: Each column reports estimates from regressions of standardized observation scores on classroom characteristics for the survey-eligible sample. Standard errors are reported in parentheses and clustered at the teacher level in specifications without teacher-by-school fixed effects; robust standard errors are reported when teacher-by-school fixed effects are included. All models include year fixed effects. Column 1 includes the classroom quality index and classroom demographics; Column 2 adds teacher characteristics (gender, race/ethnicity, age, experience, tenure, degree status, National Board Certification, and certification type); Column 3 adds school fixed effects; Column 4 includes teacher-by-school fixed effects; Columns 5–6 separately estimate models including only the classroom quality index or only classroom demographics. Both the dependent variable and the classroom quality index are standardized by year (mean = 0, SD = 1). The classroom quality index is the classroom-level average of a student index derived via principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table A.10: Effects of Classroom Characteristics on Student Survey's Indexes

	Peer support (1)	Classroom rigor (2)	Academic press (3)	Course clarity (4)	Academic engagement (5)	Academic personalism (6)	Classroom disruptions (7)
Classroom quality (std)	0.109**	0.175***	0.130***	0.123***	0.060	0.071*	0.165***
	(0.045)	(0.044)	(0.041)	(0.044)	(0.041)	(0.042)	(0.043)
Perc. male	0.333	-0.245	0.099	-0.051	0.173	0.085	-0.557**
	(0.248)	(0.225)	(0.216)	(0.230)	(0.227)	(0.219)	(0.238)
Perc. Black	-1.001**	-0.424	0.005	-0.353	-0.680*	-0.841**	-1.420***
	(0.429)	(0.414)	(0.451)	(0.451)	(0.406)	(0.405)	(0.406)
Perc. non-Hispanic non-Black	0.608	0.151	0.044	-0.023	0.084	-0.253	0.154
	(0.408)	(0.381)	(0.377)	(0.417)	(0.399)	(0.390)	(0.381)
Perc. free lunch	0.708***	0.888***	0.436	0.804***	0.795***	0.946***	0.667**
	(0.263)	(0.283)	(0.273)	(0.277)	(0.283)	(0.256)	(0.275)
Perc. special ed	0.114	0.056	-0.415	-0.162	-0.009	-0.092	-0.644**
	(0.318)	(0.303)	(0.281)	(0.300)	(0.287)	(0.281)	(0.296)
Log class size	-0.179***	-0.177**	-0.173***	-0.209***	-0.148**	-0.229***	-0.100*
	(0.060)	(0.072)	(0.060)	(0.064)	(0.060)	(0.064)	(0.061)
Teacher characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Teacher-school f.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R2	0.62	0.66	0.68	0.65	0.68	0.68	0.65
N (teachers)	4,066	4,066	4,064	4,064	4,064	4,064	4,063
N	10,988	10,989	10,985	10,980	10,981	10,982	10,982

Notes: Each column reports estimates from regressions of standardized student survey indices on classroom characteristics. Robust standard errors are reported in parentheses (teacher-by-school fixed-effects specifications). The sample includes teachers in the survey sample. All models include year fixed effects, teacher characteristics, and teacher-by-school fixed effects. Both the dependent variables and the classroom quality index are standardized by year (mean = 0, SD = 1). The classroom quality index is the classroom-level average of a student index derived via principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table A.11: Exogeneity Checks: Class-Level Correlations of Classroom Shocks

	Classroom quality (std)					
	t-1		*t*		*t+1*	
	(1)	(2)	(3)	(4)	(5)	(6)
Perc. male	-0.459***	0.060	-0.831***	-0.607***	-0.277***	0.173
	(0.092)	(0.109)	(0.065)	(0.077)	(0.086)	(0.107)
Perc. Black	-0.813***	0.066	-0.826***	-0.751***	-0.804***	0.151
	(0.019)	(0.165)	(0.015)	(0.121)	(0.019)	(0.154)
Perc. non-Hispanic non-Black	0.375***	0.161	0.343***	0.568***	0.484***	0.219
	(0.054)	(0.134)	(0.046)	(0.123)	(0.056)	(0.136)
Perc. free lunch	-1.857***	0.122	-2.066***	-0.965***	-1.900***	0.071
	(0.057)	(0.114)	(0.048)	(0.102)	(0.061)	(0.101)
Perc. special ed	-1.240***	0.081	-2.318***	-1.735***	-1.242***	0.333**
	(0.099)	(0.127)	(0.081)	(0.102)	(0.104)	(0.135)
Log class size	-0.031***	-0.027	-0.003	-0.029*	-0.043***	-0.034*
	(0.011)	(0.021)	(0.008)	(0.016)	(0.011)	(0.020)
Teacher characteristics		Yes		Yes		Yes
Teacher-school f.e.		Yes		Yes		Yes
Joint p-value	0.00	0.60	0.00	0.00	0.00	0.01
R2	0.52	0.88	0.54	0.88	0.51	0.89
N (teachers)	6,838	6,838	10,354	10,354	6,817	6,817
N	17,270	17,270	29,078	29,078	17,264	17,264

Notes: Each column reports estimates from regressions of the classroom quality index in prior, current, and future years on classroom demographic shares. Standard errors are reported in parentheses and clustered at the teacher level in specifications without teacher-by-school fixed effects; robust standard errors are reported when teacher-by-school fixed effects are included. The sample includes teachers in the observation sample. All models include year fixed effects; even-numbered columns additionally include teacher characteristics and teacher-by-school fixed effects. Both the dependent variables and the classroom quality index are standardized by year (mean = 0, SD = 1). The classroom quality index is the classroom-level average of a student index derived via principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table A.12: Sorting of Teachers to Classrooms

	Classroom quality (std)	
	(1)	(2)
Black	-0.895***	-0.066***
	(0.022)	(0.015)
Hispanic	-0.230***	-0.064***
	(0.021)	(0.013)
Other race (non-black-white-hispanic)	0.024	-0.005
	(0.042)	(0.022)
Male	-0.058**	-0.030***
	(0.023)	(0.010)
Age	-0.006***	-0.002**
	(0.001)	(0.001)
Years of experience	0.012***	0.002**
	(0.002)	(0.001)
Tenured	0.088***	0.040***
	(0.022)	(0.013)
Graduate degree	0.006	-0.006
	(0.020)	(0.011)
National Board Certification	0.108***	0.002
	(0.034)	(0.019)
No teaching certification	0.004	0.054*
	(0.043)	(0.029)
School f.e.		Yes
R2	0.15	0.64
N (teachers)	10,354	10,354
N	29,078	29,078

Notes: Each column reports estimates from regressions of the classroom quality index on teacher characteristics. Standard errors are reported in parentheses and clustered at the teacher level. The sample includes teachers in the observation sample. All models include year fixed effects; Column 2 additionally includes school fixed effects. The classroom quality index is standardized by year (mean = 0, SD = 1) and defined as the classroom-level average of a student index derived via principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table A.13: Teacher Characteristics and the Effect of Adjusting Survey Scores for Classroom Characteristics

	Two-group tournament			Three-group tournament			Quality-adjusted score		
	Survey ranking change	Exit bottom 5%	Exit top 5%	Survey ranking change	Exit bottom 5%	Exit top 5%	Survey ranking change	Exit bottom 5%	Exit top 5%
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Male	0.04	-0.03*	-0.02	0.03	-0.01	0.02	0.01	-0.00	-0.00
	(0.04)	(0.02)	(0.02)	(0.05)	(0.02)	(0.03)	(0.01)	(0.00)	(0.00)
Black	0.56***	-0.01	0.06**	0.23***	0.02	0.04	0.20***	-0.00	0.01
	(0.04)	(0.02)	(0.03)	(0.05)	(0.02)	(0.03)	(0.01)	(0.00)	(0.01)
Hispanic	0.21***	0.04	0.02	0.13**	0.03	0.00	0.04***	0.00	-0.01
	(0.05)	(0.03)	(0.03)	(0.06)	(0.03)	(0.03)	(0.01)	(0.00)	(0.01)
Other race (non-White)	0.02	-0.03**	0.01	-0.12	0.01	-0.01	-0.01	0.02	-0.00
	(0.07)	(0.01)	(0.07)	(0.08)	(0.04)	(0.07)	(0.02)	(0.02)	(0.01)
Age	0.00*	0.00	-0.00	-0.00	0.00	0.00	0.00**	0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Years of experience	-0.00	-0.00*	0.00	0.00	-0.00	0.00	-0.00**	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tenured	-0.16***	0.01	-0.01	-0.07	0.00	-0.03	-0.02*	-0.00	0.00
	(0.05)	(0.03)	(0.04)	(0.06)	(0.03)	(0.04)	(0.01)	(0.00)	(0.01)
Graduate degree	-0.02	0.00	0.01	-0.07	-0.01	-0.00	-0.00	-0.00	-0.01
	(0.04)	(0.02)	(0.03)	(0.05)	(0.02)	(0.03)	(0.01)	(0.00)	(0.01)
National Board Certification	-0.18***	0.11	0.11*	-0.12*	-0.01	0.04	-0.05***	0.00	0.07*
	(0.07)	(0.08)	(0.06)	(0.07)	(0.04)	(0.05)	(0.02)	(0.00)	(0.04)
No teaching certification	0.24	-0.08***	0.27*	0.32*	-0.05**	0.25	0.03	0.00	-0.00
	(0.15)	(0.03)	(0.15)	(0.17)	(0.02)	(0.15)	(0.04)	(0.00)	(0.00)
R2	0.03	0.07	0.06	0.01	0.02	0.03	0.05	0.09	0.08
N	10,991	557	552	10,991	557	552	10,990	557	552

Notes: Each column reports regressions of measures of adjusted survey-based teacher rankings on teacher characteristics. Standard errors are reported in parentheses and clustered at the teacher level. The sample includes teachers in the survey sample. Ranking change is the difference between a teacher's percentile ranking in the adjusted and unadjusted survey-score distributions. Exit bottom equals 1 if a teacher moves above the 5th percentile in the adjusted distribution, conditional on being below it in the unadjusted distribution. Exit top equals 1 if a teacher moves below the 95th percentile in the adjusted distribution, conditional on being above it in the unadjusted distribution. Columns 1–3 divide teachers into two groups based on whether their classroom quality index is below or above the median; Columns 4–6 divide teachers into three groups (low, medium, high); Columns 7–9 use a regression-based adjustment for classroom characteristics. The omitted racial group is White, non-Hispanic teachers. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

Table A.14: Effects of Classroom Characteristics on the Likelihood of Leaving the School

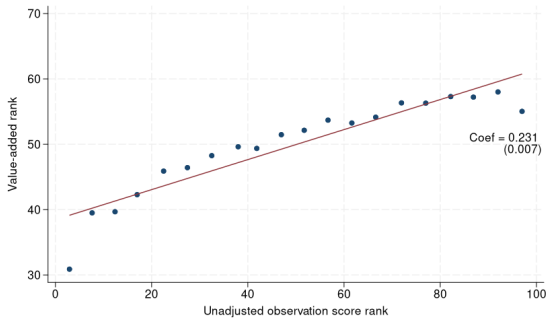
	2+ years of data sample			Untenured sample	Low rated t-1 sample
	Observation score	Leave school		Leave school	Leave school
	(1)	(2)	(3)	(4)	(5)
Classroom quality (std)	0.068***	-0.017***	0.002	0.015	-0.059
	(0.013)	(0.003)	(0.006)	(0.023)	(0.062)
Class demographics	Yes	Yes	Yes	Yes	Yes
Teacher characteristics	Yes		Yes	Yes	Yes
Teacher-school f.e.	Yes		Yes	Yes	Yes
R2	0.84	0.03	0.70	0.80	0.89
N (teachers)	10,354	10,354	10,354	4,082	1,459
N	29,078	29,078	29,078	7,385	1,944

Notes: Each column reports estimates from regressions of classroom observation or school exit on classroom characteristics. Standard errors are reported in parentheses and clustered at the teacher level in specifications without teacher-by-school fixed effects; robust standard errors are reported when teacher-by-school fixed effects are included. Columns 1–3 include teachers with at least two years of data to compute school exit in year t+1; Column 4 additionally restricts the sample to nontenured teachers; and Column 5 restricts the sample to teachers who had an "Unsatisfactory" or "Basic" rating in the prior year. All models include year fixed effects and classroom demographics, and some specifications additionally include teacher characteristics and teacher-by-school fixed effects. Both classroom observation scores and the classroom quality index are standardized by year (mean = 0, SD = 1), and leave school is an indicator for whether the teacher exits the school in $t + 1$. The classroom quality index is the classroom-level average of a student index derived via principal component analysis of baseline test scores and behavioral variables. Asterisks denote statistical significance at the ***p<0.01, **p<0.05, and *p<0.10 levels.

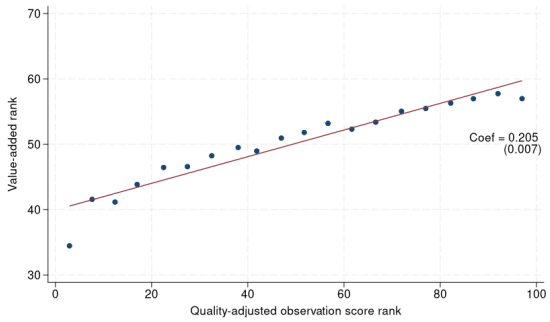
Appendix Figures

Figure A.1: Teacher Value-Added and the Effects of Adjusting Observation and Survey Scores for Classroom Composition

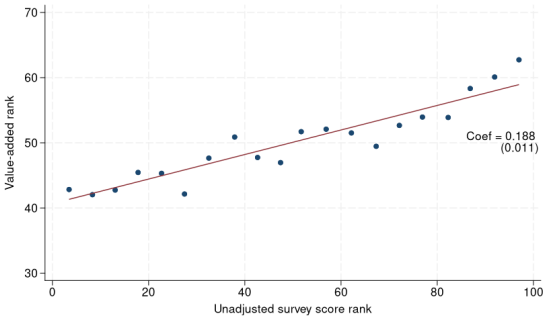
(a) Unadjusted observation score



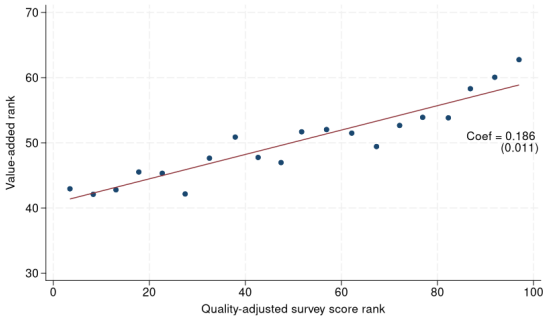
(b) Quality-adjusted observation score



(c) Unadjusted survey score



(d) Quality-adjusted survey score



Notes: Figures display binned scatter plots of teacher performance rankings (observation and survey scores) against teacher value-added rankings. Each plot divides teachers into 20 equal-sized bins based on the x-axis variable and plots the mean of the y-axis variable within each bin. The red line shows the best linear fit using unbinned data. Panels A and B display results for unadjusted and quality-adjusted observation scores, respectively; Panels C and D display corresponding results for survey scores. Teacher value-added is the average of math and reading value-added, and all rankings are computed annually.