Educational Evaluation and Policy Analysis
March 2021, Vol. 43, No. 1, pp. 60–82
DOI: 10.3102/0162373720970204
Article reuse guidelines: sagepub.com/journals-permissions

What Explains the Race Gap in Teacher Performance Ratings? Evidence From Chicago Public Schools

Matthew P. Steinberg
George Mason University

Lauren Sartain
The University of North Carolina at Chapel Hill

Racial gaps in teacher performance ratings have emerged nationwide across newly implemented educator evaluation systems. Using Chicago Public Schools data, we quantify the magnitude of the race gap in teachers' classroom observation scores, examine its determinants, and describe the potential implications for teacher diversity. Between-school differences explain most of the race gap and within-school classroom-level differences—poverty, incoming achievement, and prior-year misconduct of a teacher's students—explain the remainder of the race gap. Teachers' value-added scores explain none of the race gap. Leveraging within-teacher variation in the teacher–evaluator race match, we find that racial mismatch does not influence observation scores. Adjusting observation scores for classroom and school context will generate more equitable ratings of teacher performance and mitigate potential adverse consequences for teacher diversity.

Keywords: *classroom context, classroom observation scores, equity, teacher evaluation*

Introduction

EDUCATOR evaluation systems are holding teachers more accountable than ever before for their classroom performance. These systems not only require professional development for teachers with low evaluation ratings, but also link teacher ratings to employment termination and tenure revocation (Steinberg & Donaldson, 2016).[1] And despite efforts to incorporate multiple measures of teacher performance into these new systems, implementation challenges have limited (or even precluded) the use of student test scores to measure teacher performance.[2] As a result, classroom observations of a teacher's instructional performance continue to account for the majority (and, in some cases, the entirety) of a teacher's evaluation rating upon which high-stakes teacher personnel decisions are based (Ross & Walsh, 2019; Steinberg & Donaldson, 2016). However, while measures of teacher performance based on student achievement scores (e.g., value-added measures [VAMs]) make statistical adjustments to account for heterogeneity in the characteristics of a teacher's students, teacher ratings based on classroom observations make no such adjustments.

Yet, it has long been known that teachers are nonrandomly sorted across (and within) schools (Monk, 1987). Higher-performing teachers tend to be systematically assigned to higher-achieving and lower-poverty schools and students (Allensworth et al., 2009; Clotfelter et al., 2006; Goldhaber et al., 2015; Ingersoll, 2001; Kalogrides et al., 2013; Kalogrides & Loeb, 2013; Monk, 1987), and teachers of color tend to be assigned to schools and classrooms with lower-achieving and more economically disadvantaged students than their

White colleagues (Kalogrides et al., 2013). Given recent evidence that teachers assigned to classrooms with lower-achieving students receive lower classroom observation ratings (Campbell & Ronfeldt, 2018; Gill et al., 2016; Steinberg & Garrett, 2016; Whitehurst et al., 2014), variation in observed teacher performance may be due less to variation in the quality of a teacher's performance and more to the contextual features of their schools and classrooms. Taken together, the nonrandom patterns by which teachers are assigned to classrooms across and within schools may have differential and adverse consequences for the performance ratings of teachers based on evaluators' observations of a teacher's classroom practice. Indeed, if minority teachers are disproportionately assigned to more economically disadvantaged and lower-achieving classrooms, and the achievement of a teacher's students plays a critical role in determining teachers' observation scores, then a race gap in teacher performance ratings may emerge under newly implemented evaluation systems.

Recent evidence finds such emergent disparities in teacher evaluation ratings. In Boston Public Schools, Black and Hispanic teachers are more likely to receive low evaluation ratings than their White colleagues and therefore are more likely to be targeted for dismissal from the district (Vaznis, 2013a, 2013b). In an anonymous urban district which implemented a new educator evaluation system in the 2012–2013 school year, a disproportionally large percentage of Black teachers (relative to districtwide averages) were rated below proficient compared with their White peers, and the gap in teacher performance ratings between Black and White teachers persisted in each of the first 3 years of the district's new evaluation system (Bailey et al., 2016). In Michigan, teachers of color are more likely to receive the lowest evaluation rating compared with their same-school peers under the state's newly implemented teacher evaluation system (Drake et al., 2019).

In this article, we quantify this emergent race gap in teacher performance ratings, examine the determinants of the gap, and describe the potential equity consequences for the diversity of the teacher workforce. To do so, we examine the extent to which context-specific factors—due to classroom-level teacher sorting—and the subjective appraisals of teacher performance made by evaluators explain differences in teacher performance ratings. We address the following questions: (a) Do teacher, classroom, and school characteristics explain the observed differences in teacher performance ratings? (b) Does the teacher–evaluator race match influence teacher performance ratings? Given prior evidence that the race match between students and their teachers affects the academic achievement of students (Dee, 2004, 2005; Egalite et al., 2015; Gershenson et al., 2018), we extend this line of research by examining whether assignment to a racially similar evaluator shapes the performance ratings of teachers.

We address these questions in the context of Chicago Public Schools (CPS) and its newly implemented teacher evaluation system, Recognizing Educators Advancing Chicago's Students (REACH). We focus our analysis on the 2013–2014 and 2014–2015 school years, the first 2 full years of CPS' REACH system.[3] We leverage a unique data set that provides detailed information on elementary school (i.e., Grade K–5) teachers in Chicago, the students in their self-contained classrooms, and classroom observation-level data allowing us to match teachers to the unique evaluator responsible for observing and rating the teacher's classroom practice. By matching teachers to their observation-specific evaluator, we can examine whether variation in teacher observation ratings is due to observational differences (i.e., race) between the teacher and his or her evaluator.

Consistent with prior evidence on teacher sorting (Kalogrides et al., 2013), we find that Black teachers in Chicago are more likely than their White colleagues to teach economically disadvantaged and lower-achieving students. Moreover, Black teachers systematically teach in schools with significantly worse organizational climates—less effective leadership, fewer collaborative opportunities among teachers, less support for ambitious instruction, and weaker connections to their students' families—than their White peers. As a consequence of the nonrandom sorting of Black teachers across (and within) the most disadvantaged schools in Chicago, a large Black–White gap in teacher performance has emerged, on the order of two thirds of a standard deviation. The magnitude of this

race gap remains unchanged across the first 2 years of Chicago's REACH evaluation system, even as average observation scores improved for all teachers (and by race) between the 2013–2014 and 2014–2015 school years.

Examining the determinants of this gap, we find that the characteristics of a teacher's classroom—poverty, incoming achievement and prior-year misconduct of a teacher's students—independently explain approximately one third of the observed race gap in classroom observation scores. Teachers receive higher classroom observation scores in schools characterized by better instructional support and school leadership; yet, the organizational climate of schools explains none of the variation in the race gap, suggesting that the benefit teachers receive from teaching in more supportive learning environments does not vary by race. And, while more effective teachers—those who contribute more to student achievement growth (based on a teacher's prior-year VAM)—receive higher classroom observation scores, these quality differences explain none of the residual race gap. It is only with the inclusion of school fixed effects—which control for all observed and unobserved school-level differences—that the remaining observed race gap statistically disappears.

When we decompose the Black–White gap into three distinct sources of variability—teacher, classroom, and school levels—we find that these sources jointly explain 81% of the Black–White gap in observation scores, and the residual Black–White gap becomes no longer statistically distinguishable from zero. Between-school differences account for 89% of the explained Black–White gap and within-school differences (i.e., classroom-level differences) account for the remaining 11% of the explained Black–White gap. These findings reveal that evaluation systems which do not account for cross-school and cross-classroom differences in teachers' educational settings will generate both misleading and inaccurate ratings of teacher performance based on classroom observation scores.

Importantly, we find that variation in the teacher–evaluator race match does not influence teachers' classroom observation scores. These results rely on within-teacher variation in the teacher–evaluator race match, allowing us to control for unobserved teacher-level heterogeneity (such as teacher ability endowments) that may be correlated with teacher observation scores independent of a teacher's instructional performance during a given classroom observation. These results should mitigate concern among policymakers and school leaders in Chicago that demographic differences between teachers and their evaluators explain differences in teacher performance ratings.

Finally, we simulate the distribution of teacher ratings, by race, based on classroom observation scores that are, alternatively, unadjusted and adjusted for school and classroom characteristics. We show that a disproportionate share of Black teachers—nearly twice the rate at which they are represented among elementary school teachers in Chicago—would be ranked in the bottom quartile of the teacher performance distribution if ratings were based solely on unadjusted classroom observation scores. In contrast, we show that if classroom observation scores are adjusted for school- and classroom-level teacher sorting, the performance distribution—both the bottom and top quartiles of teacher ratings—would reflect the racial distribution of elementary school teachers in Chicago. Taken together, these results indicate that policymakers and school leaders should account for the school and classroom settings in which teachers are located to more equitably and fairly evaluate and rate teacher performance.

Conceptualizing Differences in Teacher Performance Ratings

What might explain observed differences, by race, in teacher performance ratings across educator evaluation systems? First, the classroom environment in which teachers work—in particular, the characteristics of a teacher's students—might contribute to differences in teacher ratings. Recent evidence indicates that teachers assigned to classrooms with lower-achieving students receive lower performance ratings based on classroom observations. In their analysis of four geographically distinct urban districts in the United States, Whitehurst et al. (2014) find that 29% of teachers who were assigned students with the lowest incoming achievement—the lowest quintile of the student achievement distribution—were ranked in the bottom quintile based

on their classroom observation scores; in contrast, 37% of teachers assigned the highest-performing students were ranked in the top quintile. Taking advantage of the random assignment of teachers to (classes of) students in the Measures of Effective Teaching (MET) study, Steinberg and Garrett (2016) show that the incoming achievement of a teacher's students has a significant and substantive effect on teacher performance ratings—teachers assigned to a class with one standard deviation higher incoming achievement would score one third of a standard deviation higher on their classroom observation ratings.

Second, differences in teacher ratings might be due to differences in the school-specific resources available to support instruction and teacher professional development. Using teacher and student survey data from Chicago Public Schools, researchers have identified five essential, school-based supports—leadership, professional capacity, parent–community relationships, student-centered learning climate, and ambitious instruction—that not only vary significantly across schools but have also been shown to predict significant gains in student achievement (Sebring et al., 2006). Although there is no direct evidence on whether teacher performance ratings reflect school-specific differences in these school-based supports, evidence from the experimental rollout of a teacher evaluation pilot in Chicago indicates that intensive principal instructional coaching significantly improves teacher performance and student achievement (Steinberg & Sartain, 2015a, 2015b). Given the importance of a school's learning and instructional climate for student outcomes, the extent to which these supports vary across schools might also explain differences in teacher performance ratings.

Third, differences in observation ratings may reflect variation in the distribution of teacher effectiveness across schools. In Washington state, Goldhaber et al. (2015) show that teacher quality, as measured by a teacher's value-added contribution to student achievement (VAM), is inequitably distributed across multiple indicators of student disadvantage, including poverty and achievement. Although Goldhaber et al. (2015) do not describe whether teacher quality varies by teacher background characteristics (such as race), the authors do show that higher-quality teachers (as measured by VAM) are disproportionately distributed among schools serving fewer minority students.

Finally, observable differences between teachers and evaluators might influence teacher performance ratings. Indeed, Drake et al. (2019) posit that (though, do not empirically assess whether) racial mismatch between a teacher and evaluator might introduce bias into teacher evaluation ratings. The possibility that racial mismatch might introduce bias into teacher evaluation ratings is informed by prior work by Grissom and Keiser (2011), who find that teachers who share the same race as their school principal report higher job satisfaction and are less likely to exit their school, and that Black teachers earn less than their White colleagues in the same school when their principal is White. Grissom and Keiser (2011) further find that teachers receive more encouragement, recognition, support, and individual autonomy when there is racial congruence with their school principal. Studying how the racial match between managers and employees shape employees' labor market outcomes, Giuliano et al. (2011) find that employees have better outcomes—lower quit rates, lower dismissal rates and higher promotion rates—when they share the same race as their manager. The extent to which racial congruence between teachers and evaluators influences teachers' performance ratings remains an open question, and one which we pursue in this article.

Teacher Evaluation in Chicago Public Schools

Beginning in the 2012–2013 school year, Chicago Public Schools (CPS) joined other districts across the nation by instituting sweeping reforms to its teacher evaluation system. Required to act by a new state law[4] and building on lessons learned from an earlier pilot of an evidence-based classroom observation protocol (Sartain et al., 2011; Steinberg & Sartain, 2015b), REACH was a dramatic departure from CPS's traditional system of evaluating teachers. The district's traditional, 45-year-old evaluation system relied on a checklist-based approach to observing and rating teacher practice; evidence found that this traditional evaluation system

failed to both differentiate teachers in terms of their effectiveness and provide useful feedback to improve their instructional practice (Weisburg et al., 2009).

Under the district's traditional evaluation system, teachers' summative, end-of-year ratings were based solely on one cursory annual classroom observation. Under REACH, evaluators use a detailed observation rubric to observe and rate teacher practice during multiple classroom observations in an evaluation cycle. While REACH incorporates multiple teacher performance measures—including professional practice based on classroom observation scores and student growth based on student achievement scores—into teachers' summative ratings, observation scores account for 70% to 100% of a teacher's summative evaluation score (see Table A1 in the Appendix, available in the online version of the journal).[5] The summative ratings that teachers receive under the REACH system—Unsatisfactory, Developing, Proficient or Excellent—are tied to high-stakes decisions, including tenure and dismissal.

Under REACH, teachers are observed and evaluated by their school principal or assistant principal(s), who conduct all observations of teachers in their school and are required to pass a certification examination before serving as an evaluator. Evaluators use the CPS Framework for Teaching, a modified version of the Charlotte Danielson Framework for Teaching (Chicago Public Schools, 2012; Danielson, 2011) to observe and rate a teacher's classroom practice.[6] Nontenured teachers (teachers in their first 3 years in CPS) are observed four times each year; three of these are formal observations and one can be an informal observation. Traditionally, CPS teachers earned tenured at the start of their fourth year. However, since the introduction of REACH, teacher tenure decisions are more closely tied to teachers' summative evaluation scores. Tenured teachers accumulate their four observations over 2 years, with one formal and one informal observation conducted in each school year. Notably, tenured teachers with previous low ratings (i.e., Unsatisfactory or Developing) receive four observations and a summative evaluation rating annually.

Formal observations are typically the length of a class period (approximately 45 minutes), are prescheduled, and must also include a pre- and postobservation conference between the evaluator and the teacher. For every formal classroom observation, teachers and their evaluators must also meet for pre- and postobservation conferences, during which evaluators provide teachers with detailed feedback and guidance on their instructional practice based on evaluator ratings from the classroom observation. Informal observations are unannounced, a minimum of 15 minutes and evaluators are required to give either written or in-person feedback to teachers following the observation. Both formal and informal observations are weighted equally and contribute to a teacher's summative evaluation rating, which informs teacher dismissal decisions and remediation plans. Nontenured teachers receive a summative evaluation rating annually, and most tenured teachers receive a summative evaluation rating every other year.

Dismissal, remediation, and tenure attainment policies are directly tied to a teacher's summative rating under the new REACH system. Nontenured teachers whose ratings are in the bottom two rating categories (i.e., Unsatisfactory and Developing) may not have their contracts renewed. Tenured teachers with a Developing rating are placed on a Professional Development Plan, which remains in effect for 1 year. Tenured teachers with an Unsatisfactory rating are subject to a 90-day Remediation Plan and subject to dismissal if their ratings do not improve. Notably, summative REACH ratings affect the order in which teachers are laid off—with lower rated teachers being the first to be dismissed (Chicago Public Schools, 2016).

Data and Sample

We employ administrative data on teachers and their evaluators from the 2013–2014 and 2014–2015 school years, the first 2 years of REACH. For each teacher, we observe demographic information (race, gender, age), experience (years teaching in CPS), degree attainment (master's degree or higher), and tenure status. We match teachers to the unique evaluator who conducted and scored each of the teacher's classroom observations. The matching of teachers to evaluators across multiple classroom observations enables us to construct a teacher ×

observation × year data set for the first 2 full years of REACH. For each evaluator, we observe his or her demographic information (race, gender, age), experience (years spent in current position [i.e., principal or assistant principal] in current school), and the evaluator's formal role (principal or assistant principal). We also rely on student-level administrative data to match classes of students to their teachers. Below, we describe the teacher and evaluator samples, the classroom- and school-specific characteristics which capture the context in which teachers do their work, and the measure of teacher performance based on classroom observation scores.

Teacher and Evaluator Samples

In CPS, elementary school teachers (i.e., Grade K–5 teachers) teach either self-contained classes or multiple classes within (or across) grades. We focus on CPS teachers who teach in self-contained classrooms—those teachers who teach multiple subjects to the same class of students throughout the school day. By focusing on teachers in self-contained classrooms, we aim to avoid the concern that teacher performance ratings, which depend on multiple classroom observations of a teacher's instructional performance, may be based on different groups of students.

Table 1 summarizes the characteristics of Grade K–5 teachers teaching in self-contained classrooms. In the 2013–2014 and 2014–2015 school years, 5,536 K–5 teachers taught in 411 CPS elementary schools. Among K–5 teachers, 93% are female, 42% are White, 23% are Black and 28% are Latino; 76% are tenured in the district and 65% have a master's degree. These K–5 teachers have, on average, 10.8 years of experience in the district. In addition to the full sample of K–5 teachers in self-contained classrooms, we construct two subsamples: (a) the VAM sample includes K–5 teachers for whom a value-added measure (VAM) from the prior school year is available; and (b) the Race Match sample includes K–5 teachers evaluated at least twice in a given school year and who had at least one evaluator of the same race and at least one evaluator of a different race.[7] K–5 teachers in Chicago share very similar observable characteristics with CPS teachers districtwide, with the exception of gender (93% female compared with 77% female among all CPS teachers) and a larger share of Latino teachers (28% compared with 20% among all CPS teachers).

Using matched teacher–evaluator data at the observation level, we construct a sample of evaluators responsible for observing and scoring a teacher's instructional practice (see Table A2 in the online version of the journal). Among evaluators who observed and rated K–5 teachers teaching in self-contained classrooms in Chicago, 50% are school principals, with 3.8 years of experience, on average, in their current school and in their current position (i.e., principal or assistant principal). The majority of evaluators are female (71%), 31% of evaluators are White, 44% are Black, and 21% are Latino. Evaluators responsible for rating the instructional practice of K–5 teachers look similar to all school administrators in Chicago responsible for evaluating teacher performance. Of the 21,912 unique observations at the teacher × observation × year level in the K–5 sample, 51% of observations were conducted by principals and the remaining 49% were conducted by assistant principals.

Classroom Characteristics

By matching teachers to their self-contained classroom of students, we can examine the extent to which within-school sorting of teachers to classes explains observed differences in teacher performance ratings. Specific attention is paid to two characteristics of a teacher's students that have elsewhere been found to explain differences in teacher performance based on classroom observation scores—student achievement and poverty (Steinberg & Garrett, 2016; Whitehurst et al., 2014). We use the incoming (i.e., prior year) achievement of a teacher's students on end-of-year standardized exams (from the reading portion of the NWEA exam), which we standardize at the subject × grade × year level. Classroom-level poverty is based on the percentage of a teacher's students receiving free- or reduced-price lunch (FRPL). Teacher performance ratings may also be shaped by the behavioral characteristics of the class. We therefore include a third classroom characteristic—the incoming (i.e., prior year) misconduct record of the teacher's students.[8] To do so, we rely on detailed student-level infraction data to construct

TABLE 1

Teacher Characteristics

Teacher characteristics	District	K–5 teachers	VAM	Race match
Age	40.7 (11.1)	39.6 (10.8)	41.0 (10.4)	38.3 (10.8)
Female	0.77	0.93	0.89	0.94
White	0.51	0.42	0.45	0.49
Black	0.22	0.23	0.26	0.15
Latino	0.20	0.28	0.20	0.33
Other race	0.07	0.07	0.08	0.03
Experience	11.0 (7.9)	10.8 (7.5)	11.8 (7.1)	9.9 (7.3)
Tenured	0.73	0.76	0.85	0.69
Master's degree	0.69	0.65	0.69	0.63
Teacher-year observations	41,833	8,633	2,697	1,360
Teachers	22,068	5,536	1,847	1,119
Schools	531	411	371	157

Note. Data are pooled from the 2013–2014 and 2014–2015 school years. Proportions are reported, except for age and experience, which report mean (standard deviation) in years. *K–5 Teachers* includes classroom teachers in Grade K–5 teaching multiple subjects to the same class of students (i.e., self-contained classrooms). *VAM* is a subset of the *K–5 Teachers* and includes teachers for whom a value-added measure (VAM) from the prior school year is available. *Race Match* is a subset of the *K–5 Teachers* and includes teachers evaluated at least twice in a given school year and who had at least one evaluator of the same race and at least one evaluator of a different race. *Experience* is the number of years of teaching experience in Chicago Public Schools (CPS). For the 47 tenured teachers missing experience data, we impute the sample mean for tenured teachers of 13.31 years; for the 133 nontenured teachers missing experience data, we impute a value of 2 years.

a measure of incoming student behavior; specifically, we count the number of prior-year misconducts committed by each of a teacher's students, and normalize the total count on a per-capita basis based on the current year's student enrollment in a teacher's class.

School Climate

In addition to the characteristics of students in a teacher's classroom, the extent of school-specific resources available to support instruction and teacher professional development might also explain differences in teacher performance ratings. Indeed, prior evidence from Chicago finds that a framework of five essential supports and contextual resources are necessary for school improvement efforts and are correlated with student achievement growth (Sebring et al., 2006). We incorporate school-level data on the five measures identified by the UChicago Consortium which constitute these essential supports, including (a) leadership, (b) professional capacity, (c) parent–community relationships, (d) student-centered learning climate, and (e) ambitious instruction. Using school-level data on these five measures, we construct an index of school climate as an equal weighting of the five survey measures which we standardize (mean zero, standard deviation one) at the school × year level (see Table A3 in the online version of the journal for results from a principal components analysis [PCA] which identifies one underlying component describing school climate based on the five measures). We note that the school climate measure varies only at the school × year level and not within school × year cells (i.e., there is no variation in the school climate measure at the school × grade × year level).

Table 2 summarizes the classroom characteristics and school climate to which K–5 teachers—both overall and by race—are exposed to in Chicago. Here, we report classroom-level differences among teachers in the VAM sample for whom data are available on the incoming (i.e., prior-year) academic achievement of their students (patterns of classroom-level poverty, misconduct, and school climate among all K–5 teachers in Chicago are nearly identical to the VAM sample of teachers; see Table 2). Consistent with prior evidence on teacher sorting (Kalogrides et al., 2013), Black teachers in

Chicago are more likely than their White colleagues to teach economically disadvantaged and lower-achieving students. Black teachers teach in classrooms where 94% of students, on average, receive FRPL (i.e., classroom poverty); this compares to 80% of students, on average, who receive FRPL in White teachers' classrooms (Table 2, Panel B). Students in Black teachers' classrooms score, on average, 0.23 standard deviations below the districtwide mean in reading; in comparison, students in White teachers' classrooms score 0.14 standard deviations above the districtwide mean. Students in Black teachers' classrooms also have significantly higher rates of misconduct; relative to their White peers, students taught by Black teachers have more than twice the number of (prior-year) behavioral misconducts—5.84 misconducts per-pupil, on average—compared with 2.61 misconducts per-pupil, on average, among students in White teachers' classrooms. Moreover, Black teachers systematically teach in schools with significantly worse organizational climates than their White peers. Based on our school climate measure, Black teachers teach in schools that are 0.25 standard deviations below the districtwide mean; White teachers teach in schools that are 0.13 standard deviations above the districtwide mean.

Teacher Performance Ratings: Classroom Observation Scores

Despite significant attention in recent years to the use of student test scores in the construction of teacher performance measures (i.e., VAMs), approximately three-quarters of all teachers nationwide teach in nontested grades and/or subjects where student-test-score data—on which VAM scores are based—are unavailable (Watson et al., 2009; Whitehurst et al., 2014). In contrast, all teachers nationwide are evaluated based on classroom observations of their instructional performance. Our measure of teacher performance relies on observations of a teacher's classroom practice. Under REACH, nontenured teachers receive four observations during each school year, and tenured teachers receive two observations during each school year. For each classroom observation, we observe a teacher's scores on each of nine components of practice. The nine components capture one of two domains of classroom practice—the Classroom Environment (Domain 2) or Instruction (Domain 3)—based on CPS's Framework for Teaching (FFT), a modified version of Charlotte Danielson's Framework for Teaching classroom observation protocol.[9] For each of nine components, teachers receive a score from 1 to 4 on an integer scale, with 1 indicating *Unsatisfactory*, 2 indicating *Developing*, 3 indicating *Proficient*, and 4 indicating *Distinguished*.

Following recent empirical applications which measure teacher performance using classroom observation scores (Garrett & Steinberg, 2015; Kane et al., 2013; Mihaly et al., 2013), we construct a lesson-specific observation score by averaging across the nine FFT components at the teacher level.[10] For analyses examining the influence of school- and classroom-level factors on teacher performance ratings, we create a summative, year-specific score by averaging across multiple lesson-specific observation scores from both formal and informal classroom observations. For analyses assessing whether racial mismatch between teachers and evaluators influences teachers' performance ratings, the outcome is a teacher's lesson-specific observation score.[11]

Table 3 summarizes teacher performance ratings based on teachers' classroom observation scores, both overall and by teacher race. Two important patterns emerge. First, classroom observation scores, on average, improve for all teachers (and by race) between the 2013–2014 and 2014–2015 school years. Second, there is a substantively large and statistically significant gap in observation scores between Black teachers and their White peers—0.63 standard deviations—and this race gap persists and remains unchanged across the first 2 years of Chicago's REACH evaluation system. The Latino–White race gap is more modest in magnitude (0.10 standard deviations).

Empirical Approach

We first assess the role that teacher, classroom, and school factors may play in explaining differences in observation scores by teacher race. To do so, we examine the extent to which differences in teacher scores, by race, may be explained by teacher-level sorting across schools and across

TABLE 2

Classroom and School Characteristics, by Teacher Race/Ethnicity

	All teachers	White	Black	Latino	Other race
Panel A: *K–5 Teachers*					
Classroom poverty	0.86 (0.22)	0.79 (0.27)	0.93 (0.14)	0.92 (0.16)	0.82 (0.23)
Classroom misconduct	1.92 (4.88)	1.63 (4.42)	3.46 (6.72)	1.14 (3.44)	1.74 (4.12)
School climate	0.04 (1.02)	0.18 (1.02)	−0.29 (0.90)	0.10 (1.03)	0.03 (1.11)
Teachers	5,536	2,363	1,302	1,495	376
Schools	411	389	288	266	211
Panel B: *VAM*					
Classroom poverty	0.86 (0.22)	0.80 (0.27)	0.94 (0.12)	0.90 (0.17)	0.85 (0.20)
Classroom misconduct	3.36 (6.40)	2.61 (5.70)	5.84 (8.38)	2.10 (4.09)	2.72 (5.19)
Classroom achievement	−0.02 (0.58)	0.14 (0.58)	−0.23 (0.50)	−0.14 (0.55)	0.05 (0.57)
School climate	0.01 (1.00)	0.13 (1.01)	−0.25 (0.90)	0.07 (0.99)	−0.01 (1.05)
Teachers	1,847	818	485	396	148
Schools	371	297	200	158	108

Note. Data are pooled from the 2013–2014 and 2014–2015 school years. Each cell reports teacher-level mean (standard deviation). *Classroom Poverty* is measured as the proportion of a teacher's students who receive free/reduced-price lunch. *Classroom Misconduct* is measured as the per-capita count, based on the student enrollment of a teacher's class, of prior-year student behavioral infractions. *Classroom Achievement* is measured as the incoming (i.e., prior year) academic achievement of a teacher's students on the NWEA reading exam (standardized at the subject × grade × year level) and reported in standard deviation units. *School Climate* is a school-level index of school-based supports and contextual resources for school improvement, and is measured as an index of five survey measures based on teacher and student surveys conducted by the UChicago Consortium; the five measures include leadership, professional capacity, parent–community relationships, student-centered learning climate, and ambitious instruction. The *School Climate* index is constructed as an equal weighting of the five survey measures (see Table A3 in the online version of the journal for results from a principal components analysis) and standardized (mean zero, standard deviation 1) at the school × year level. Panel A (*K–5 Teachers*) includes classroom teachers in Grade K–5 teaching multiple subjects to the same class of students (i.e., self-contained classrooms); Panel B (*VAM*) is a subset of the *K–5 Teachers* and includes teachers for whom a value-added measure (VAM) from the prior school year is available.

TABLE 3

Teacher Observation Scores, by Teacher Race

	Observation score			Difference in observation score (vs. White teachers)		
	2013–2014	2014–2015	Pooled	2013–2014	2014–2015	Pooled
All teachers	3.07 (0.48)	3.13 (0.47)	3.10 (0.48)			
White	3.16 (0.46)	3.22 (0.46)	3.19 (0.46)			
Black	2.86 (0.47)	2.92 (0.48)	2.89 (0.48)	−0.64***	−0.63***	−0.63***
Latino	3.11 (0.45)	3.17 (0.42)	3.14 (0.44)	−0.09**	−0.11***	−0.10***
Other race	3.08 (0.48)	3.15 (0.48)	3.11 (0.48)	−0.17***	−0.16**	−0.16***
Teacher-year observations	—	—	8,633			
Teachers	4,279	4,354	5,536			

Note. Observation Score reports mean (standard deviation) of teachers' unadjusted classroom observation scores in FFT points. A teacher's unadjusted observation score is the average of formal and informal classroom observation scores, which are based on the Chicago Framework for Teaching, a modified version of the Danielson Framework for Teacher (FFT) classroom observation protocol, and includes nine components across two domains of classroom practice—Classroom Environment (Domain 2) and Instruction (Domain 3). Each component is rated on a 1 to 4 integer scale. *Difference in Observation Score (vs. White Teachers)* reports difference in observation score (by teacher race) relative to White teachers and is reported in standard deviation units. Standard deviation units are statistically significant at the *10%, **5%, and ***1% levels. Sample includes teachers in the *K–5 Teachers* sample.

classrooms within the same school. We estimate variants of the following full model:

$$Score_{jst} = \beta_0 + \boldsymbol{Race}_j`\zeta + \boldsymbol{Classroom}_{jst}`\delta + \pi VAM_{js,t-1} + \rho Climate_{jst} + X_{jt}`\Gamma + \lambda_t + \theta_s + \gamma_g + \upsilon_{jst},$$

where *Score* is teacher *j*'s summative observation score based on formal and informal observations in school *s* during school year *t*. $Race_j$ is a full set of teacher-race dummies, with White as the omitted reference category. *Classroom* is a vector of classroom-level characteristics of teacher *j*'s students in school *s* during year *t* that we show vary by teacher race, including classroom poverty, the incoming academic achievement of a teacher's students, and the incoming misconduct record of a teacher's students (see Table 2). *Climate* is the school climate index to which teacher *j* is exposed to in school *s* during school year *t*, and *VAM* is teacher *j*'s value-added score from the prior school year. Teacher *j*'s VAM score is the average of (where available) math and reading VAM scores from school year $t - 1$ and is the individual-level VAM score incorporated into a teacher's summative evaluation rating (see Table A1 in the online version of the journal). X is a vector of observable teacher characteristics, including gender, age, tenure status, and academic attainment (i.e., master's degree or not). The terms λ_t, θ_s, and γ_g represent year, school and grade fixed effects, respectively, and υ_{jst} is a random error term. We cluster the standard errors at the school level. In alternative models, we replace the summative observation score with domain-specific scores which capture teacher performance based on either the teacher's management of the classroom environment (FFT Domain 2) or the teacher's delivery of instruction (FFT Domain 3).

Next, we examine whether variation in teacher observation scores may be explained by observed differences between teachers and their evaluators. That is, to what extent might teachers' observation scores vary as a function of racial mismatch between teachers and their evaluators. To do so, we rely on within-year, within-classroom (i.e., within-teacher) variation in teacher–evaluator matches to estimate the influence of racial mismatch on teacher performance ratings. Our strategy relies on the fact that for some teachers, observations of different classroom lessons are conducted by different evaluators. This particular feature of the classroom observation process in CPS allows for a teacher fixed effects approach, enabling us to account for any influence that fixed, unobserved teacher characteristics (including a teacher's endowed instructional ability) may have on measured performance. Furthermore, the fact that the fixed effects estimates are generated from multiple observations conducted within the same school year mitigates concerns that changes in teacher practice and effectiveness across multiple school years may bias these estimates (Table 1 summarizes teacher characteristics for teachers evaluated at least twice in a given school year and who had at least one evaluator of the same race and at least one evaluator of a different race; Table A2 in the online version of the journal summarizes evaluator characteristics).

To estimate the influence of racial mismatch in the teacher–evaluator match, we estimate variants of the following model:

$$Score_{jlt} = \beta_0 + \beta_1 (OtherRace_{jlt}) + X_{jt}`\Gamma + Z_{jlt}`\zeta + \theta_{jt} + \varepsilon_{jlt},$$

where *Score* is teacher *j*'s classroom observation score for lesson *l* in school year *t* (teacher scores from both formal and informal classroom observations are included). *OtherRace* is a binary indicator variable which equals 1 if the race of teacher *j* and the evaluator of lesson *l* in school year *t* differ, and zero otherwise. X is a vector of observable teacher characteristics, including gender, age, tenure status, and academic attainment (i.e., master's degree or not). Z is a vector of observable characteristics of the evaluator of teacher *j* during lesson *l* in year *t*, including gender, age, and an indicator for the evaluator's formal role (principal or assistant principal). The term θ_{jt} is a teacher × year fixed effect and ε_{jlt} is a random error term. This approach enables a within-teacher comparison across multiple lessons observed in the same school year. We cluster the standard errors at the school level.

Results

School and Classroom Context

What explains the large and persistent race gap in teacher performance ratings? Is the race

gap due to differences in the classroom environment in which teachers teach? Variation in the school-specific resources available to support instruction and teacher professional development? Differences in how teacher effectiveness is distributed across schools? Are there residual differences in teacher performance ratings even after accounting for teacher-, classroom-, and school-level factors?

Table 4 summarizes evidence on these questions. First, we find that neither observable teacher characteristics nor a teacher's effectiveness at improving student achievement, as measured by a teacher's prior-year VAM score, explain any of the gap—0.32 observation points, or 0.63 standard deviations—in teacher performance ratings between Black and White teachers (Table 4, columns 2 and 3). This is despite the fact that more effective teachers, on average, receive higher classroom observations ratings. Indeed, teachers with one standard deviation higher prior-year VAM are rated 0.13 points (or 0.27 standard deviations) higher, on average, on their classroom observations (Table 4, column 3).

Evidence that a teacher's prior-year VAM score does not explain any of the gap in observation scores may also reflect the relatively weak correlation between VAM and classroom observation scores. Prior evidence across multiple studies finds that the correlation between VAM and classroom observation scores is on the order of .10 to .30 (Jiang et al., 2014; Kane & Staiger, 2012; Steinberg & Kraft, 2017). In our VAM sample, the correlation between prior-year VAM and current-year observation score is .24 (in the K–5 sample, the correlation between prior-year VAM and current-year observation score is 0.25).

In contrast, the inclusion of observable dimensions of teachers' classrooms reduces the Black–White gap in observation ratings from −0.32 to −0.22 observation points (Table 4, column 4). This result reflects the disproportionate assignment of Black teachers to same-grade classrooms (i.e., grade fixed effects) across schools serving the most academically and economically disadvantaged students in Chicago. Independently, the extent of poverty in a teacher's classroom and the incoming misconduct and achievement levels of a teacher's students each significantly predict teacher observation scores across same-grade teachers. These results indicate that teachers teaching in more economically and academically disadvantaged classes receive lower observation scores, on average. Notably, a one standard deviation increase in incoming student achievement is associated with a 0.10 point increase in measured teacher performance; the magnitude of this relationship is consistent with experimental evidence showing that the incoming achievement of a teacher's students affects teacher performance based on classroom observation scores (Steinberg & Garrett, 2016).[12]

Furthermore, teachers receive higher observation ratings, on average, when teaching in better school climates—schools characterized by better instructional leadership and more resources to support ambitious instruction and teacher professional development, as well as more supportive relationships with students' parents. Teachers teaching in schools with one standard deviation better organizational climates are rated 0.05 points (or 0.10 standard deviations) higher, on average, on their classroom observations (Table 4, column 5). Yet, the organizational climate of schools explains none of the variation in the race gap, suggesting that the benefit that teachers receive from teaching in more supportive learning environments does not vary by race.

It is only after controlling for all observable and, importantly, unobservable school-level differences that the residual race gap in teacher ratings between Black and White teachers becomes statistically indistinguishable from zero (Table 4, column 6). Notably, the incoming achievement of a teacher's students and a teacher's own contribution to student achievement growth, even within the same schools and among teachers in the same grade-level, remain an important determinant of teacher ratings (Table 4, column 6). These results point to significant within-school variation in both classroom composition and teacher effectiveness in Chicago, a result that has been documented elsewhere (see, e.g., Goldhaber et al., 2015).[13]

Figure 1 visually illustrates the role that school- and classroom-level factors play in explaining race-specific differences in teacher performance ratings. We present the distribution of teacher observation scores, by race, that are adjusted for the same set of school- and classroom-level factors as in our main regression results (Table 4). In Panel A, we find significant

TABLE 4

Estimated Difference in Teacher Observation Scores, by Teacher Race/Ethnicity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Black	−0.32*** (.033)	−0.32*** (.032)	−0.32*** (.030)	−0.22*** (.030)	−0.21*** (.031)	−0.06 (.035)	−0.06 (.039)
Latino	−0.03 (.027)	−0.04 (.027)	−0.03 (.027)	0.02 (.026)	0.01 (.025)	0.004 (.022)	0.01 (.025)
Other race	−0.09** (.038)	−0.11*** (.037)	−0.10*** (.035)	−0.07** (.034)	−0.07** (.034)	−0.04 (.030)	−0.04 (.033)
VAM (prior year)			0.13*** (.014)	0.11*** (.013)	0.11*** (.012)	0.13*** (.014)	0.14*** (.016)
Classroom poverty				−0.24*** (.078)	−0.18** (.075)	−0.21* (.129)	−0.13 (.145)
Classroom misconduct				−0.007*** (.002)	−0.006*** (.002)	−0.002 (.001)	−0.001 (.002)
Classroom achievement				0.10*** (.022)	0.10*** (.022)	0.09*** (.024)	0.11*** (.026)
School climate					0.05*** (.016)	−0.02 (.020)	
P value from Black = Latino	.000	.000	.000	.000	.000	.112	.125
Teacher × year observations	2,697	2,697	2,697	2,697	2,697	2,697	2,697
Teachers	1,847	1,847	1,847	1,847	1,847	1,847	1,847
Schools	371	371	371	371	371	371	371
Adjusted R^2	0.091	0.131	0.177	0.245	0.255	0.473	0.486
Teacher characteristics		X	X	X	X	X	X
Grade FE				X	X	X	X
School FE						X	
School × year FE							X

Note. Coefficients reported with robust standard errors (clustered at the school level) in parentheses. Data are for the 2013–2014 and 2014–2015 school years. Sample includes teachers in the *VAM* teacher sample. The dependent variable is a teacher's unadjusted classroom observation score (in FFT points). All regressions control for year fixed effects. *Other Race* includes teachers who did not identify as White, Black, or Latino. Teacher characteristics include age, gender, experience, tenure status, and educational attainment (i.e., master's degree). *VAM (prior year)* is a teacher's value-added score from the prior academic year. *Classroom Poverty* is the proportion of a teacher's students who receive free/reduced-price lunch. *Classroom Misconduct* is the per-capita count of prior-year student behavioral infractions among a teacher's students. *Classroom Achievement* is the incoming (i.e., prior year) academic achievement of a teacher's students and reported in standard deviation units. *School Climate* is a school-level index of school-based supports and contextual resources for school improvement and is reported in standard deviation units. Coefficients statistically significant at the *10%, **5% and ***1% levels. FFT = Framework for Teaching; VAM = value-added measure.

race-specific heterogeneity in teacher performance based on unadjusted observation scores. As shown in Table 4, the distribution of race-specific observation scores are unaffected by the inclusion of both observable teacher characteristics (Panel B) and teacher effectiveness (Panel C). Yet, after adjusting for classroom characteristics (poverty, achievement, misconduct), the race-specific distributions of teacher performance ratings are much more similar (Panel D). And, with the full set of classroom- and school-specific controls (see Table 4, column 6), the distribution of teacher performance differs little by teacher race (Panel F).

Finally, we decompose the Black–White gap into three distinct sources of variability: teacher, classroom, and school levels. Figure 2 presents the proportion of the Black–White gap in observation scores that is explained by these sources of variability, both independently and jointly (see Table A6 in the online version of the journal for the coefficient estimates upon which Figure 2 is based). We find that teacher-specific differences, including teacher characteristics and a teacher's prior-year VAM score, explain none of the Black–White gap in observation scores, while classroom-level differences independently explain 31% of the Black–White gap and school-level differences independently explain 72% of the Black–White gap. Together, teacher, classroom, and school factors explain 81% of the Black–White gap in observation scores, though the residual Black–White gap (−.06 points) is not statistically distinguishable from zero (see Table 4 [column 7] and Table 5 [column *Between & Within School*]). These findings indicate that 89% of the explained Black–White gap is due to between-school differences (i.e., 0.72/0.81 from Figure 2) while 11% of the explained Black–White gap is due to within-school differences (i.e., 0.09/0.81 from Figure 2).[14]

Teacher–Evaluator Race Match

To what extent does the racial mismatch between teachers and evaluators explain teacher performance ratings based on classroom observation scores? Table 5 presents these results. First, relying just on cross-school variation in teacher observation scores, we find that the performance of Black and White teachers is substantively and significantly associated with having an other-race evaluator (Table 5 column 1), and this association is unchanged with the inclusion of both observable teacher characteristics (column 2) and observable evaluator characteristics (column 3). Specifically, controlling for observable characteristics of teachers and evaluators, White teachers score 0.16 points (or 0.33 standard deviations) lower on their observation scores when their evaluator is of a different race, compared with White teachers' observation scores from same-race evaluators. In contrast, Black teachers score 0.11 points (or 0.23 standard deviations) higher on their observation scores when their evaluator is of a different race compared to Black teachers' observation scores from same-race evaluators. Given that evaluators, like teachers, are also sorted to schools in nonrandom ways, White teachers who have an other-race evaluator are more likely to teach in more economically and academically disadvantaged schools; in contrast, Black teachers who have an other-race evaluator are more likely to teach in more advantaged schools. However, once we condition on school fixed effects (Table 5, column 4), these associations effectively disappear in magnitude and become statistically insignificant. These patterns hold when teacher performance is evaluated in terms of either a teacher's performance in managing the classroom environment (Table A7 in the online version of the journal) or a teacher's instructional performance (Table A8 in the online version of the journal).

Next, we leverage multiple classroom observations conducted in the same school year for the same teacher (i.e., teacher × year fixed effects) to estimate the influence of an other-race evaluator on teacher observation scores (Table 5, column 5). We find no evidence that differences in the teacher–evaluator race match influences teacher performance ratings. Indeed, conditional on teacher × year fixed effects, racial mismatch has no relationship with a teacher's overall performance rating (Table 5, column 5), ratings of a teacher's performance in managing the classroom environment (Table A7, column 5), or ratings of a teacher's instructional performance (Table A8, column 5). We further find no evidence of race-specific heterogeneity in the role of racial mismatch. Indeed, other-race evaluators

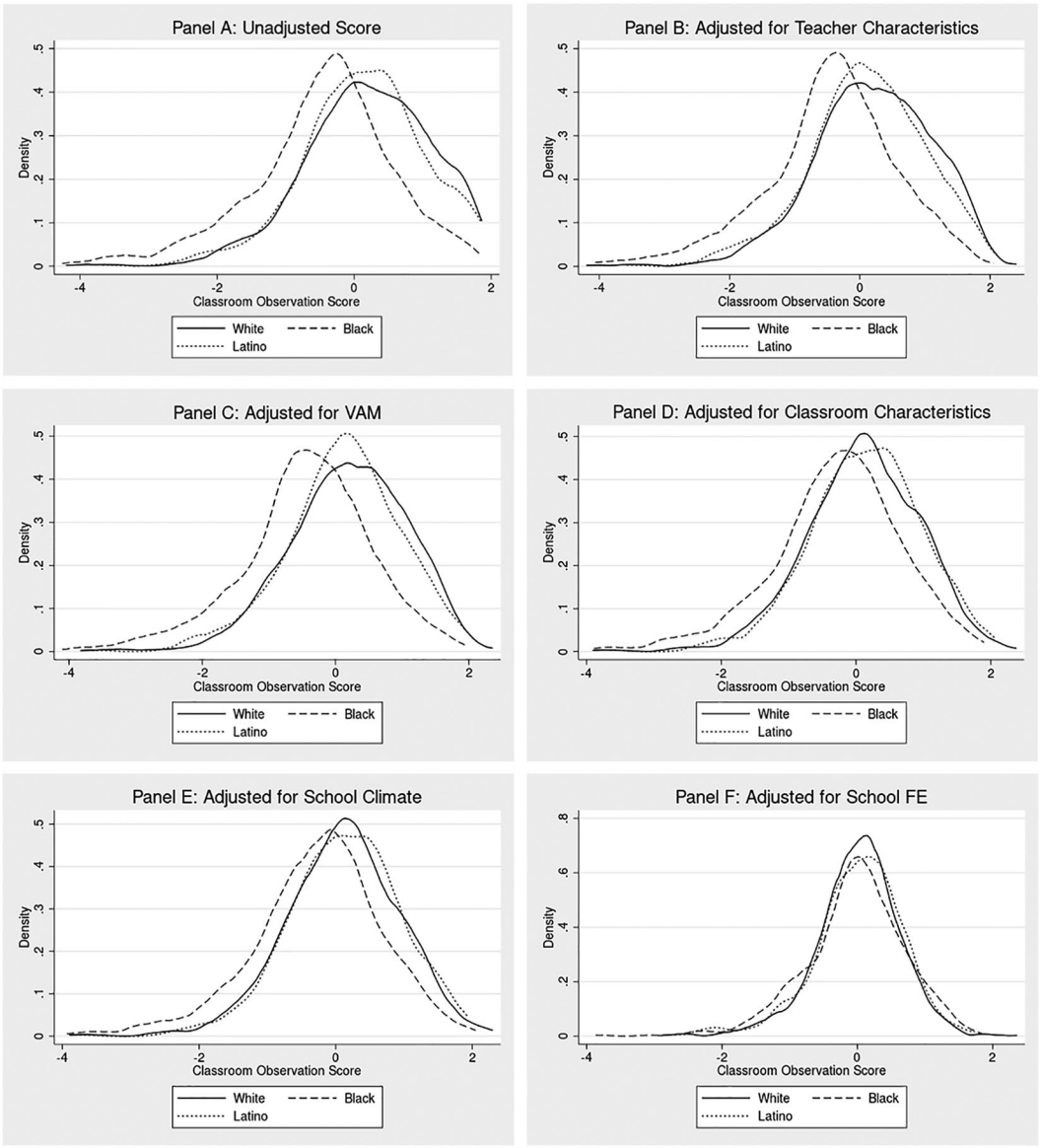

FIGURE 1. *Distribution of teacher observation scores, by teacher race.*
Note. Data are from the 2013–2014 and 2014–2015 school years. Panel A shows the distribution of unadjusted teacher observation scores, by teacher race (Table 4, column 1). Panel B shows the distribution of teacher observation scores, by teacher race, adding controls for teacher characteristics (Table 4, column 2). Panel C shows the distribution of teacher observation scores, by teacher race, adding controls for teacher's prior-year VAM score (Table 4, column 3). Panel D shows the distribution of teacher observation scores, by teacher race, adding controls for classroom characteristics (Table 4, column 4). Panel E shows the distribution of teacher observation scores, by teacher race, adding controls for school climate (Table 4, column 5). Panel F shows the distribution of teacher observation scores, by teacher race, adding controls for school fixed effects (Table 4, column 6). See Table 4 for more detail on the teacher, classroom, and school-level characteristics included as controls. VAM = value-added measure.

do not differentially rate the performance of White, Black, or Latino teachers (see Table 5, column 5, Panels B–D).[15]

The teacher × year fixed effects approach enables us to account for unobserved, time-invariant teacher and classroom-level heterogeneity that may be correlated with a teacher's classroom observation scores. However, other-race evaluators were not randomly assigned across teachers nor across

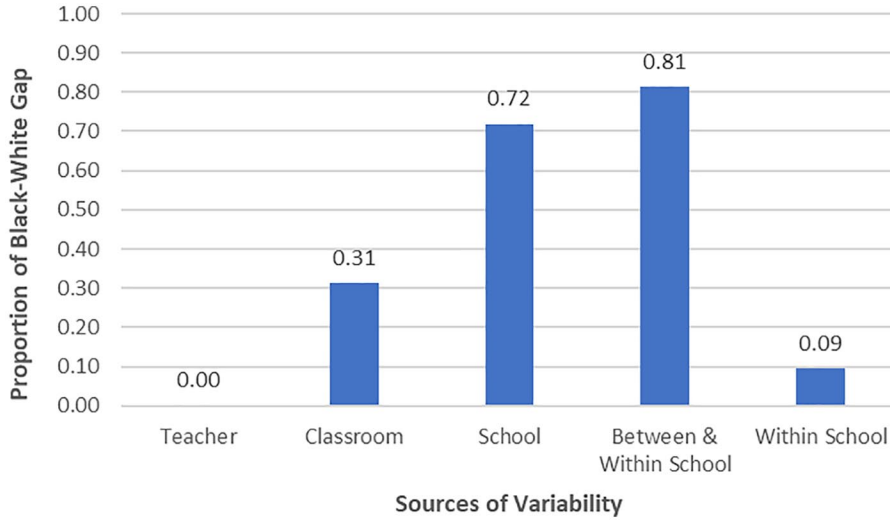

FIGURE 2. *Proportion of the Black–White gap explained, by sources of variability.*
Note. Each vertical bar indicates the proportion of the unadjusted Black–White gap in teacher observation scores explained by teacher, classroom, and/or school factors (see Table A6 in the online version of the journal). The proportion of the unadjusted race gap explained is calculated as one minus the ratio of the estimated coefficient on the race variable (i.e., Black, from Equation 1) to the unadjusted race gap (from Table 4, column 1). *Within-School* is calculated by subtracting the proportion of the gap explained by *School* factors from the proportion of the gap explained by *Between & Within School* factors. Table A6 reports the coefficient estimates upon which the values of the vertical bars are calculated.

TABLE 5

Teacher–Evaluator Race Match and Teacher Observation Scores

	(1)	(2)	(3)	(4)	(5)
Panel A: All teachers					
Other-race evaluator	0.01 (.016)	0.01 (.016)	0.01 (.016)	−0.01 (.012)	−0.01 (.015)
Teachers	5,346	5,346	5,346	5,346	5,346
Evaluators	932	932	932	932	932
Schools	411	411	411	411	411
Panel B: White teachers					
Other-race evaluator	−0.17*** (.029)	−0.16*** (.029)	−0.16*** (.029)	0.00 (.023)	0.00 (.025)
Teachers	2,285	2,285	2,285	2,285	2,285
Evaluators	852	852	852	852	852
Schools	388	388	388	388	388
Panel C: Black teachers					
Other-race evaluator	0.11*** (.038)	0.11*** (.038)	0.11*** (.039)	−0.03 (.034)	−0.01 (.038)
Teachers	1,250	1,250	1,250	1,250	1,250
Evaluators	598	598	598	598	598
Schools	284	284	284	284	284
Panel D: Latino teachers					
Other-race evaluator	−0.01 (.033)	−0.01 (.032)	−0.00 (.033)	−0.03 (.029)	−0.02 (.037)
Teachers	1,446	1,446	1,446	1,446	1,446
Evaluators	568	568	568	568	568
Schools	262	262	262	262	262

(continued)

TABLE 5 (CONTINUED)

	(1)	(2)	(3)	(4)	(5)
Teacher characteristics		X	X	X	X
Evaluator characteristics			X	X	X
School FE				X	
Teacher × year FE					X

Note. Coefficients reported with robust standard errors (clustered at the school level) in parentheses. Data are for the 2013–2014 and 2014–2015 school years. The dependent variable is a teacher's unadjusted classroom observation scores (in FFT points). Columns (1) to (4) include year fixed effects. Teacher characteristics include age, gender, experience, tenure status, and master's degree attainment. Evaluator characteristics include age, gender, and an indicator for the evaluator's formal role (principal or assistant principal). In Panel A, the count of teachers is less than the count of teachers in the *K–5 Teachers* sample because we drop teachers for whom their evaluators had missing race data. There are 21,912 unique observations at the teacher × observation × year level in the K–5 sample, of which 3,738 observations (17.1%) are included in the Race Match sample. Among the 3,738 unique observations at the teacher × observation × year level in the Race Match sample, 46% of observations were conducted by principals and the remaining 54% were conducted by assistant principals. Coefficients statistically significant at the *10%, **5%, and ***1% levels.

observations for the same teacher. As a result, time-varying unobserved shocks to teachers and/or their classrooms that are correlated with both the assignment of an other-race evaluator and teacher performance could introduce bias into the teacher × year fixed effects estimates. For example, if student and/or teacher performance improves over the course of the school year, and other-race evaluators are systematically less likely (than same-race evaluators) to evaluate classroom lessons occurring later in the year, then our estimates would (incorrectly) reveal differences in observation ratings due to racial mismatch. Although we are unable to account for all unobserved, time-varying factors that may introduce bias into the teacher × year fixed effects estimates, we examine whether the assignment of other-race evaluators varies across multiple annual classroom observations among teachers evaluated by both same- and other-race evaluators within a school year.

Table 6 summarizes the distribution of other-race evaluators across multiple annual observations, by teacher race. Panel A presents results for nontenured teachers receiving four annual observations; Panel B presents results for tenured teachers receiving two annual observations. Results from a chi-square test of independence indicate that the distribution of other-race evaluators across observations (within a school year) does not vary by teacher race, both for nontenured and tenured teachers. These results should help mitigate concerns that the patterns of racial mismatch—that is, the timing of when teachers were observed by evaluators of the same or different race—might bias the teacher fixed effects estimates.

Implications for Teacher Diversity

What are the potential consequences for the diversity of the teacher workforce when performance ratings are based on classroom observation scores that do not account for the school and classroom settings in which teachers of different races are disproportionately assigned? To explore this issue, we simulate the distribution of teacher ratings, by race, based on classroom observation ratings that are, alternatively, unadjusted and adjusted for school and classroom characteristics.

Table 7 summarizes results which report the proportion of teachers, by race, in the bottom quartile of the performance distribution based on the rank order of observation ratings (Panel A) and the proportion of teachers, by race, in the top quartile of the performance distribution (Panel B). First, we present the raw distribution of teacher race. Among the VAM sample, 49% of teachers are White, 29% are Black, and 22% are Latino (all other-race teachers are excluded from this analysis due to small sample sizes). Yet, when classroom observation scores are unadjusted for the classroom and school settings in which teachers work—as is current practice in teacher evaluation systems like REACH and other systems nationally—a disproportionate share of Black teachers would be ranked in the

TABLE 6

Distribution of Other-Race Evaluator and Observation Order, by Teacher Race/Ethnicity

	All teachers	White	Black	Latino
Panel A: Nontenured teachers				
Observation 1	0.21	0.19	0.30	0.20
Observation 2	0.28	0.27	0.27	0.30
Observation 3	0.26	0.25	0.30	0.25
Observation 4	0.25	0.28	0.13	0.25
Teachers	176	100	29	47
Observations by other-race evaluator	340	191	60	89
Panel B: Tenured teachers				
Observation 1	0.48	0.48	0.57	0.47
Observation 2	0.52	0.52	0.43	0.53
Teachers	407	181	51	175
Observations by other-race evaluator	407	181	51	175

Note. Each cell reports the proportion of observations conducted by an other-race evaluator. Observation 1 is the first classroom observation in a school year; Observation 2 is the second classroom observation in a school year; Observation 3 is the third classroom observation in a school year; and Observation 4 is the fourth classroom observation in a school year. Data on the order of classroom observations only available in the 2014–2015 school year. The analysis is restricted to teachers in the *Race Match* sample in 2014–2015; teachers receiving more or less than the required annual observations are excluded. Panel A includes nontenured teachers who received the required four annual classroom observations; Panel B includes tenured teachers who received the required two annual classroom observations. In Panel A, the χ^2 statistic for test of independence is 1.393 ($p = .238$); in Panel B, the χ^2 statistic for test of independence is 5.118 ($p = .163$).

bottom quartile of the teacher performance distribution. Specifically, while 38% of the lowest rated teachers would be White, 46% of the lowest rated teachers would be Black, nearly twice the rate of the raw distribution of teacher race in our sample (Table 7, Panel A). In contrast, 64% of the highest rated teachers would be White and just 12% of the highest rated teachers would be Black based on unadjusted observation scores (Table 7, Panel B).

However, when classroom observation scores are adjusted for school- and classroom-level teacher sorting, the distribution of teacher performance by race—both the bottom and top quartiles of teacher ratings—reflects the racial distribution of elementary school teachers in Chicago. Specifically, 47% of the lowest rated teachers would be White and 32% would be Black if observation scores were adjusted for classroom- and school-level characteristics (see *+School FE* column, Panel A). Similarly, 47% of the highest rated teachers would be White and 29% of the highest rated teachers would be Black if observation scores were adjusted for classroom- and school-level characteristics (see *+School FE* column, Panel B).

Discussion

Evidence from Chicago that the race gap in teachers' classroom observation scores reflects differences in the school and classroom settings in which teachers teach, rather than real differences in teacher performance, should be of concern to policymakers and school leaders. Not only does the existence of a race gap in teacher ratings lead to a misleading and inaccurate ranking of teacher performance, but this inaccurate ranking may also have real implications for the diversity of the teacher workforce under newly implemented evaluation systems. Indeed, when classroom observation scores do not account for the school and classroom settings in which teachers teach, we find that a disproportionate share of Black teachers are ranked in the lowest quartile of teacher performance. As a result, Black teachers may be disproportionately (and incorrectly) targeted for remediation and dismissal, relative to their White peers.

Racial disproportionality in teacher ratings, and the potential labor market consequences for minority teachers, is particularly concerning in light of both the widening demographic and racial gap between teachers and their students

TABLE 7

Simulated Distribution of the Ranking of Teacher Observation Scores, by Teacher Race/Ethnicity

	VAM sample	Unadjusted score	+ Teacher characteristics	+ VAM (prior year)	+ Classroom characteristics	+ School climate	+ School FE
Panel A: Bottom quartile of teacher scores							
White	0.49	0.38	0.36	0.36	0.42	0.42	0.47
Black	0.29	0.46	0.47	0.48	0.42	0.40	0.32
Latino	0.22	0.16	0.17	0.17	0.16	0.18	0.21
Teacher × Year observations	2,484	621	621	621	621	621	621
Panel B: Top quartile of teacher scores							
White	0.49	0.64	0.63	0.64	0.56	0.55	0.47
Black	0.29	0.12	0.13	0.12	0.18	0.18	0.29
Latino	0.22	0.24	0.24	0.24	0.27	0.27	0.24
Teacher × Year observations	2,484	617	621	621	621	621	621

Note. Each cell (within a panel) reports a proportion. Data are from the 2013–2014 and 2014–2015 school years and are based on the *VAM* sample (excluding teachers who do not identify as either White, Black, or Latino). The *VAM Sample* column reports the proportion of teachers, by race. In Panel A, columns report the proportion of teachers, by race, in the bottom quartile of the performance distribution based on the rank order of estimated residuals; in Panel B, columns report the proportion of teachers, by race, in the top quartile of the performance distribution based on the rank order of estimated residuals. The *Unadjusted* column is based on the distribution of residuals from a model that controls only for year fixed effects (Table 4, column 1). The + *Teacher Characteristics* column is based on the distribution of residuals from a model that adds controls for teacher characteristics (Table 4, column 2). The + *VAM (prior year)* column is based on the distribution of residuals from a model that adds controls for a teacher's prior-year VAM score (Table 4, column 3). The + *Classroom Characteristics* column is based on the distribution of residuals from a model that adds controls for the characteristics of a teacher's classroom, including poverty, misconduct, and incoming academic achievement (see Table 4, column 4). The + *School Climate* column is based on the distribution of residuals from a model that adds controls for the climate of a teacher's school (see Table 4, column 5). The + *School FE* column is based on the distribution of residuals from a model that adds controls for school fixed effects (see Table 4, column 6). Other-race teachers are excluded from this analysis due to small sample sizes.

and evidence that minority students realize both short- and long-term benefits to their educational experiences when exposed to minority teachers (Dee, 2004, 2005; Egalite et al., 2015; Gershenson et al., 2018; Lindsay & Hart, 2017).[16] Indeed, policymakers and school leaders should encourage the type of teacher sorting that increases opportunities for minority students to be exposed to minority teachers. That this type of nonrandom teacher sorting may be penalized by an evaluation system that does not account for heterogeneity in the characteristics of a teacher's students, however, may limit the extent to which minority teachers seek out teaching assignments in some of our nation's most economically and racially segregated schools.

These findings have important implications for both the equitable implementation of newly reformed teacher evaluation systems as well as for the academic experience of students. First, if high-stakes human capital decisions rely on observation scores that do not account for context-specific factors at the school and classroom levels, then districts run the risk of making personnel decisions that have the consequence of reducing racial diversity among their teacher labor force. Such reductions in teacher diversity will likely also affect the educational experiences of students, given the benefits to student achievement of racially similar teachers.[17] Second, if demographic differences between teachers and evaluators affect a teacher's observation scores, then the ability of newly implemented evaluation systems to fairly evaluate teacher performance and equitably make high-stakes accountability decisions, decisions that are based overwhelmingly (if not entirely) on classroom observation scores, may be called into question. Notably, however, evidence that teachers' classroom observation scores are not related to racial mismatch between teachers and their evaluators should relieve some concern that racial bias might explain differences in teacher performance ratings.

Ultimately, we have shown that teachers' observation scores are not comparable across and within schools due to the nonrandom sorting of teachers. Yet, critics of adjusting teachers' classroom observation scores for school context might argue that controlling for school-level differences implicitly excuses schools for less supportive working conditions for teachers and poorer learning conditions for students. Moreover, some may be concerned that adjusting teachers' observation scores for the demographic characteristics of their students is akin to conditioning VAM scores on student background characteristics (e.g., poverty status), with the implicit assumption that we should expect different outcomes for students of different backgrounds. However, we have shown that, in the absence of adjusting scores for classroom and school factors such as the incoming achievement of a teacher's students, high-stakes evaluation systems risk penalizing teachers—especially teachers of color—for working in more challenging school environments.

Therefore, this article offers guidance to school leaders on ways to better distinguish a teacher's effectiveness from the students in his or her classroom. This work should also inform district policymakers and school leaders on the potential implications that newly implemented evaluation systems may have on teacher diversity. Indeed, systems which rely on potentially biased measures of teacher performance will have important consequences for the diversity of the teacher workforce and for a district's ability to recruit and support teachers who represent the students they teach and the community in which its schools are located. Although our findings do not explicitly link teacher ratings based on classroom observation scores to labor market outcomes (e.g., tenure and dismissal), we show that race-specific variation in teacher performance ratings reflect how teachers are sorted across schools and provide guidance for policymakers to refine new systems of personnel management and evaluation to better account for such factors. In doing so, greater equity may be achieved in the process of evaluating teacher performance with the potential to improve the educational circumstances of teachers and students in some of Chicago's (and the nation's) most disadvantaged schools.

Acknowledgments

The authors thank Elaine Allensworth, Eric Taylor, three anonymous reviewers, and participants at the Association for Public Policy Analysis and Management (APPAM) and Association for Education Finance and Policy (AEFP) annual conferences for helpful comments and suggestions, Chicago Public

Schools for continued data access and support, and Jennie Jiang for research assistance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Lauren Sartain received funding from the Spencer Foundation in support of this research.

Notes

1. As of the 2016–2017 school year, 92% of states and 88% of the largest 25 school districts and the District of Columbia have revised their systems for evaluating teacher performance. Under these new evaluation systems, teacher performance may be based on student test scores, standards-based observations of a teacher's classroom practice, and student perception surveys (Steinberg & Donaldson, 2016). Among states with newly implemented evaluation systems, 83% link teacher summative ratings to required professional development for low-rated teachers; among the largest districts and the District of Columbia (DC), 74% require professional development for low-rated teachers. Furthermore, 61% of newly implemented state evaluation systems (and 39% of newly implemented systems in the largest school districts and DC) tie teacher ratings to employment termination, while 48% of new state systems and 22% of new district systems tie teacher ratings to tenure granting/revocation decisions (Steinberg & Donaldson, 2016).

2. Among the implementation challenges are the availability of measures of student achievement growth for teachers in nontested grades and subjects (Gill et al., 2013), negative perceptions among teachers about the use of student achievement measures to estimate student growth (Sporte et al., 2013) and resistance to the inclusion of student achievement scores in teacher evaluation systems by various stakeholders (Polikoff & Porter, 2014). More recently (i.e., over the past three school years), some states and the District of Columbia have modified their evaluation system reforms by limiting (or even removing) the requirement that student test scores be incorporated into teacher evaluations (Ross & Walsh, 2019).

3. REACH was implemented for nontenured teachers in 2012–2013 and for all teachers (both nontenured and tenured) beginning in the 2013–2014 school year.

4. The Illinois Performance Evaluation Reform Act mandates reforming systems of evaluating teacher performance.

5. REACH includes two measures of student growth: (a) value-added scores based on student achievement data and (b) student growth based on performance tasks. Performance tasks are district-created assessments and are administered and scored by teachers at the beginning and end of the school year. Only teachers who teach reading or math in Grades 3 to 8 receive an individual value-added score. Elementary teachers in nontested grades/subjects receive a school-wide literacy value-added score. Both individual and school-wide value-added scores are calculated using the NWEA MAP assessment. Previous analyses of student growth based on performance tasks have found little cross-teacher variation, with almost all teachers scoring highly on this measure of student growth (Jiang et al., 2014). See Table A1in the online version of the journal for details on the nominal weights that each teacher performance measure contributes to a teacher's summative rating. See Jiang et al. (2014) for more information on teacher performance measures based on student growth (including teacher performance tasks and individual and school-wide value-added scores).

6. The rubric for classroom teachers is the same for all grades and subjects, including those who teach special education students or English language learners. Different observation rubrics are used to observe and evaluate librarians, counselors, and education support specialists.

7. A teacher's prior-year VAM score is constructed as the average of a teacher's math VAM and reading VAM scores from the prior school year, and is the individual-level VAM score (see Table A1 in the online version of the journal) that is incorporated into a teacher's summative evaluation rating as part of Chicago's REACH system (see Jiang et al., 2014, for more detail on the construction of individual teacher VAM scores incorporated into REACH summative evaluation ratings).

8. Misconduct data include student behavioral infractions defined by the CPS student code of conduct as Level 1 (inappropriate behaviors), Level 2 (disruptive behaviors), Level 3 (seriously disruptive behaviors), Level 4 (very seriously disruptive behaviors), Level 5 (most seriously disruptive behaviors), and Level 6 (illegal and most seriously disruptive behaviors).

9. There are four components that capture a teacher's performance in managing the classroom environment (Domain 2), including (a) creating an environment of respect and rapport, (b) establishing a culture for learning, (c) managing classroom procedures, and (d) managing student behavior. There are five components that capture a teacher's instructional performance (Domain 3), including (a) communicating with students, (b) using questioning and discussion

techniques, (c) engaging students in learning, (d) using assessment in instruction, and (e) demonstrating flexibility and responsiveness (i.e., responding to students' individual learning needs).

10. We also pursued a measurement-based approach—principal components analysis (PCA)—as an alternative way of constructing teacher performance scores from classroom observations. We found one principal component where each (of the nine) FFT component received approximately equal weight in constructing the principal component (PCA results are available upon request).

11. Formal and informal observation scores contribute to a teacher's summative observation score in Chicago for high-stakes personnel decisions, and are weighted equally. Since 76% of CPS teachers in the K–5 teacher sample are tenured (and 85% of teachers in the VAM sample are tenured), most teachers will have just one formal observation score in a school year. We include both formal and informal observation scores in analyses assessing the role of teacher–evaluator matches, which allows us to increase the teacher × year sample and enables more robust within-teacher estimates of the influence of teacher–evaluator match on observation scores.

12. Steinberg and Garrett (2016) show that, among a sample of Grade 4 to 8 teachers, a teacher assigned students with one standard deviation higher incoming reading achievement would realize an increase of 0.11 observation points, equivalent to one third of a standard deviation in teacher performance. The magnitude of the effect was smaller for student math achievement; specifically, a teacher assigned students with one standard deviation higher incoming math achievement would realize an increase of 0.06 observation points, equivalent to approximately one fifth of a standard deviation in teacher performance.

13. In alternative regressions, we separately examine each of the two domains of classroom practice—managing the classroom environment (Table A4 in the online version of the journal) and instruction (Table A5 in the online version of the journal)—that constitute a teacher's overall classroom observation rating. The results indicate that the classroom- and school-specific factors similarly explain differences in teacher performance ratings, by race, for these two domains of practice. Furthermore, the conditional associations with the observation scores for each domain of practice are nearly identical across the two domains. Yet, while the Black–White gap in teacher scores for managing the classroom environment statistically disappears with the inclusion of school- and classroom-level covariates, there remains a residual Black–White gap in teacher scores for instruction; although this residual Black–White gap remains statistically significant, the magnitude of the gap—on the order of 0.08 points or .17 standard deviations—is less than 25% the size of the initial Black–White gap in teacher instruction.

14. Nearly all of the explained variation in the Black–White gap in observation scores is due to just four factors. Specifically, in alternative regressions (available upon request), we find that 75% of the Black–White gap in observation scores may be explained by classroom context (measured by incoming student achievement, prior-year student misconduct and grade fixed effects), and school-level differences (measured by school × year fixed effects).

15. The minimum detectable effect size of the teacher × year fixed effects estimates (from Equation 2) may be calculated as follows: $SE\ (\hat{\beta}_1) \times 1.96 = (.015) \times (1.96) = .0294$ FFT points of the classroom observation score (see Table 5, Panel A, column 5 for the standard error [*SE*] of the estimated coefficient on *Other-Race Evaluator* [i.e., $\hat{\beta}_1$]). Given that the standard deviation of *Score* from Equation 2 is 0.48 (see Table 3), the minimum detectable effect size is $\frac{.0294}{0.48} = .061$ standard deviations of the classroom observation score.

16. Nationally, students of color made up nearly half of all public school students in the 2011 year, while teachers of color made up only 18% of all teachers (Boser, 2014).

17. For example, Dee (2005) finds that the teachers to whom students are assigned—and, specifically, the racial match between students and teachers—have large effects on teacher perceptions of student performance. In other work, Dee (2004) finds that assignment to an own-race teacher significantly increased the math and reading achievement of both Black and White students.

References

Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). *The schools teachers leave: Teacher mobility in Chicago Public Schools*. Consortium on Chicago School Research at the University of Chicago.

Bailey, J., Bocala, C., Shakman, K., & Zweig, J. (2016). *Teacher demographics and evaluation: A descriptive study in a large urban district* (REL 2017–189). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. http://ies.ed.gov/ncee/edlabs

Boser, U. (2014, May). *Teacher diversity revisited: A state-by-state analysis*. Center for American Progress.

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

Chicago Public Schools. (2012). *2012 CPS framework for teaching*. https://www.cps.edu/globalassets/cps-pages/careers/school-leadership/principal-quality/principal-eligibility/frameworkforteaching.pdf

Chicago Public Schools. (2016). https://www.ctulocal1.org/wp-content/uploads/2018/08/CTU_Contract_2015-2019.pdf

Clotfelter, C., Ladd, H., & Vigdor, J. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources*, *41*(4), 778–820.

Danielson, C. (2011). *Enhancing professional practice. A framework for teaching* (3rd ed.). Association for Supervision and Curriculum Development.

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. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review*, *95*(2), 158–165.

Drake, S., Auletto, A., & Cowen, J. (2019). Grading teachers: Race and gender differences in low evaluation ratings and teacher employment outcomes. *American Educational Research Journal*, *56*(5), 1800–1833.

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

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*(2), 224–242.

Gershenson, S., Hart, C., Hyman, J., Lindsay, C., & Papageorge, N. (2018). *The long-run impacts of same-race teachers* (NBER Working Paper No. 25254). https://www.nber.org/papers/w25254

Gill, B., Bruch, J., & Booker, K. (2013). *Using alternative student growth measures for evaluating teacher performance: What the literature says* (REL 2013–002). U.S. Department of Education, Institute of Educational Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midatlantic. http://ies.ed.gov/ncee/edlabs

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments* (REL 2017–191). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory MidAtlantic. http://ies.ed.gov/ncee/edlabs

Giuliano, L., Levine, D., & Leonard, J. (2011). Racial bias in the manager-employee relationship: An analysis of quits, dismissals, and promotions at a large retail firm. *Journal of Human Resources*, *46*(1), 26–52.

Goldhaber, D., Lavery, L., & Theobold, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher*, *44*(5), 293–307.

Grissom, J., & Keiser, L. (2011). A supervisor like me: Race, representation, and the satisfaction and turnover decisions of public sector employees. *Journal of Policy Analysis and Management*, *30*(3), 557–580.

Ingersoll, R. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*, 499–534.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2014). *Analytic memo: Evaluation data from the first year of REACH*. University of Chicago Consortium on Chicago School Research.

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

Kalogrides, D., Loeb, S., & Beteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*(2), 103–123.

Kane, T. J., McCaffrey, D., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (Bill & Melinda Gates Foundation MET Project research paper). Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. MET Project. http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Lindsay, C., & Hart, C. (2017). Exposure to same-race teachers and student disciplinary outcomes for black students in North Carolina. *Educational Evaluation and Policy Analysis*, *39*(3), 485–510.

Mihaly, K., McCaffrey, D. F., Staiger, D., & Lockwood, J. R. (2013). *A composite estimator of effective teaching* [Technical report for the Measures of Effective Teaching project]. Bill & Melinda Gates Foundation.

Monk, D. H. (1987). Assigning elementary pupils to their teachers. *The Elementary School Journal*, *88*(2), 166–187.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, *36*(4), 399–416.

Ross, E., & Walsh, K. (2019). *State of the states 2019: Teacher and principal evaluation policy*. National Council of Teacher Quality.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. University of Chicago Consortium on Chicago School Research.

Sebring, P., Allensworth, E., Byrk, A., Easton, J., & Luppescu, S. (2006). *The essential supports for school improvement*. University of Chicago Consortium on Chicago School Research.

Sporte, S. E., Stevens, W. D., Healey, K., Jiang, J., & Hart, H. (2013). *Teacher evaluation in practice: Implementing Chicago's REACH students*. University of Chicago Consortium on Chicago School Research.

Steinberg, M. P., & Donaldson, M. (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., & 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.

Steinberg, M. P., & Kraft, M. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher*, *46*(7), 378–396.

Steinberg, M. P., & Sartain, L. (2015a). Does better observation make better teachers? *Education Next*, *15*(1), 70–76.

Steinberg, M. P., & Sartain, L. (2015b). Does teacher evaluation improve school performance? Experimental evidence from Chicago's excellence in teaching project. *Education Finance and Policy*, *10*(4), 535–572.

Vaznis, J. (2013a, April 24). Union says teacher evaluation plan has race bias. *The Boston Globe*. http://www.bostonglobe.com/metro/2013/04/23/boston-union-officials-black-and-hispanic-teachers-disproportionately-targeted-under-new-evaluation-system/LCghntHAh8zM2R8qPmYrzM/story.html

Vaznis, J. (2013b, May 24). Boston teachers receive high ratings. *The Boston Globe*. https://www.bostonglobe.com/metro/2013/05/23/boston-teachers-receive-high-ratings-stirring-concern-about-rigor-evaluations/t2yk8sb2qhtlULAJwYOa0M/story.html

Watson, J. G., S. B. Kraemer, & C. A. Thorn. 2009. *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades*. Washington, DC: Center for Educator Compensation Reform, U.S. Department of Education.

Weisburg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. The New Teacher Project.

Whitehurst, G., Chingos, M., & Lindquist, K. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Brown Center on Education Policy at Brookings.

Authors

MATTHEW P. STEINBERG is an associate professor of education policy and director of EdPolicy*Forward*: The Center for Education Policy at George Mason University. His research focuses on teacher evaluation and human capital, school discipline and safety, urban school reform, and school finance. He may be contacted at msteinb6@gmu.edu.

LAUREN SARTAIN is an assistant professor in the School of Education at the University of North Carolina at Chapel Hill and an affiliated researcher at the UChicago Consortium on School Research. She studies a range of topics in urban education policy, including teacher quality, school choice and school quality, and discipline reform. She may be contacted at lsartain@unc.edu.

Manuscript received February 11, 2020
First revision received April 1, 2020
Second revision received June 16, 2020
Third revision received September 18, 2020
Accepted September 22, 2020