Investigating Race and Gender Biases in High-Stakes Teacher Observations

Jason A. Grissom

Vanderbilt University

Brendan Bartanen

Texas A&M University

February 2020

Abstract

Classroom observation ratings make up the largest component of summary ratings given to teachers in the multiple-measure teacher evaluation systems states have implemented in the last decade, but little research has examined observation ratings in a high-stakes setting. Using data from the first six years of statewide implementation of teacher evaluation in Tennessee, we investigate whether nonrandom sorting of students and teachers and other potential sources of bias systematically lower the observation scores of teachers according to their race and gender. We find that white and female teachers outscore their Black and male colleagues, even when comparing teachers with otherwise similar characteristics in the same school with similar value-added scores. These gaps appear across rubric domains. The Black-white gap is largest in schools where Black teachers are racially isolated, and we find evidence that teachers receive somewhat higher ratings from raters of the same race. In contrast, we find no same-gender rater effects, and are in fact able to explain little of the gender gap with other observable factors.

Investigating Race and Gender Biases in High-Stakes Teacher Observations

Introduction

The widespread implementation of multiple-measure teacher evaluation systems has been a defining feature of the last decade of education reform (Steinberg & Garrett, 2016). Such systems typically pair scores from classroom observations conducted by a trained rater using a standards-based rubric with value-added or other measures of student test score growth, sometimes alongside other indicators of teacher effectiveness, such as student surveys. Multiple-measure evaluation has the potential to provide teachers both with specific feedback on the strengths and weaknesses of their classroom practices and with measures of their impacts on their students—a powerful combination, in theory, for helping teachers identify what may be working in their classrooms and what areas may need attention. Beyond these developmental purposes, however, evaluation scores have high stakes, as principals and school system leaders can also use evaluation results to inform hiring, placement, compensation, and retention or dismissal decisions.

Given both the developmental and high-stakes purposes of teacher evaluation, an important question for research and policy is whether measures generated by these evaluation systems are biased—that is, whether teachers' scores are systematically influenced by factors outside their control. A relatively large literature has examined bias in teacher value-added metrics (e.g., Chetty, Friedman, & Rockoff, 2014; Kane & Staiger, 2008; Rothstein, 2009). A chief concern in this literature is nonrandom sorting of students to teachers, such that more able students are assigned to more effective teachers, though from their review of the strongest evidence in this literature, Koedel, Mihaly, and Rockoff (2015) conclude that bias in models with sufficient adjustments for student background likely is minimal.

Much less research has examined bias in teacher observations, which comprise the largest component of the overall evaluation in most systems (Grissom & Youngs, 2016). Researchers have identified this gap as a large one in the literature, pointing out numerous

potential sources of bias in observation scores that have not been investigated (Cohen & Goldhaber, 2016). These include bias associated with the features of the rating instrument and bias introduced by raters themselves. One source that is of similar concern to those in the value-added literature is the potential for bias from nonrandom sorting. Teachers who are assigned lower-performing students or students with higher propensities toward discipline problems, for example, may be marked lower if raters do not account for student-driven differences in the classroom environments they observe. In their study of four districts, Whitehurst, Chingos, and Lindquist (2014) find that teachers assigned students with lower incoming achievement levels indeed received substantially lower ratings than teachers whose classrooms had higher-achieving students. Steinberg and Garrett (2016) find similar patterns for classrooms from the Measures of Effective Teaching (MET) project, including in a sample for which students were randomly assigned to teachers, suggesting that bias—and not just lower instructional quality in classrooms with lower-achieving students—at least partially drives ratings differences by incoming student achievement across classrooms.

We may be particularly concerned about whether evaluation scores are biased with respect to race/ethnicity and gender. Women and teachers of color can be more likely to be assigned lower-achieving students (Kalogrides, Loeb, & Béteille, 2013), so they may be more susceptible to negative biases from student sorting than male and white teachers. Other sources of bias, particularly for observation ratings, may produce different directional predictions for gender and race/ethnicity. For example, rubrics describing the characteristics of effective instruction may be based on white, female archetypes of good teaching (e.g., Salazar, 2018), which may advantage women but disadvantage teachers of color. Beyond basic issues of fairness, the degree to which biased evaluation ratings may affect teachers' work attitudes and decisions to turn over, or may subject them to personnel action, raises important policy questions at a time when increasing the racial/ethnic and gender diversity of the teacher workforce has become relevant for states

and districts (Albert Shanker Institute, 2015; U.S. Department of Education, 2016).

The small body of rigorous research on racial/ethnic and gender bias in observation ratings has been limited primarily to low-stakes settings, such as Campbell and Ronfeldt's (2018) study of MET data that found evidence that both male and Black teachers received lower ratings, with the latter finding fully explained by the composition of classrooms to which Black teachers were assigned. Yet raters' behavior in low- and high-stakes settings can differ substantially (Grissom & Loeb, 2017), and research has not yet examined whether the patterns in observation scores Campbell and Ronfeldt (2018) document extend to the case in which evaluation scores can be used for personnel decisions. An exception is the recent work of S. Drake, Auletto, and Cowen (2019), who examine summative ratings assigned to teachers in Michigan during the initial years of that state's implementation of statewide evaluation. As the authors note, however, ratings at that time were relatively unregulated, with no common expectations for classroom observations (including that they occurred) or incorporation of other measures, such as student achievement. Local determination of evaluation procedures without standardized rubrics or guidelines for how raters assign ratings more closely resembles typical state systems prior to the evaluation reform wave of the last decade (Steinberg & Garrett, 2016) and presents a very different case than the one we investigate in this study.

We examine bias in classroom observation scores in Tennessee over the first six years of the implementation of the state's high-stakes, multiple-measure teacher evaluation system (2011–12 to 2016–17). Tennessee's system required specific standards-based rubrics and procedures for teacher observation conducted by trained raters, and placed observation scores alongside test score-based measures of student achievement in a specified formula to determine teachers' overall evaluation rating. We ask specifically whether observation scores are systematically associated with teacher gender or race, and investigate factors that may contribute to such differences. An advantage of our data is that we can observe indicator-level ratings for individual observations throughout the school year (i.e., not just

average observation ratings), and we can link these observations to information about the rater who assigned the rating, students taught by the focal teacher, subject taught, and other characteristics of the school environment. For teachers in tested grades and subjects, we can also access measures of instructional quality as captured by their value-added to student test scores. Our analysis is based on data from approximately 360,000 teacher-by-year observations.

We uncover large gaps in classroom observation ratings by both teacher gender and race. Pooling across years, women outscore men by 0.32 SD and white teachers outscore Black teachers by 0.15 SD, on average. Descriptively, these gaps generally persist across school levels, locale types, observation rubrics, subjects taught, and teacher experience, though they vary in magnitude. The Black-white gap, for instance, is largest in town/rural schools and smallest in urban schools, and is approximately twice as large in high schools as in elementary schools. When we model observation scores in a regression framework, we find little evidence that differences in teacher qualifications, school characteristics, or teacher value-added explain average gender and race gaps. The Black-white gap can be explained to some degree by differences in classroom context—within schools, Black teachers tend to be assigned larger numbers of low-achieving students with higher rates of absences and disciplinary infractions, and these characteristics are linked to lower observation ratings. Moreover, leveraging variation within school and school year in the characteristics of raters (which can vary because both principals and assistant principals conduct classroom observations), we find that teachers receive higher scores when they have a same-race rater, which increases the Black-white gap because white teachers are more likely to be race-matched. In contrast, our estimate of the gender gap is remarkably consistent across models, regardless of other factors we include in our models; we can do little to explain why men persistently score lower than women. Moreover, we find no evidence that teachers benefit from being observed by a rater of the same gender.

An important finding, however, is that substantial heterogeneity exists in the

Black-white gap by school context. Isolating within-school comparisons of Black and white teachers, we show that bias against Black teachers is largest when they are racially isolated, while the racial gap in observation scores disappears in schools that have a majority of Black teachers. This pattern holds even after we account for the increased presence of Black students and administrators in schools with many Black teachers.

Potential Sources of Bias in Teacher Evaluation Scores

In the wake of Race to the Top (RTTT), 46 states reformed their teacher evaluation processes to collect more clearly defined measures of teacher effectiveness (Steinberg & Garrett, 2016). Traditional systems based exclusively on infrequent classroom observations using broad performance checklists typically produced undifferentiated ratings in which nearly all teachers were deemed effective. Post-RTTT evaluation reforms put standards-based observation rubrics into widespread use and paired them with measures of teacher effectiveness based on student test score growth, and sometimes with other metrics, such as feedback from student surveys (Grissom & Youngs, 2016). Scores produced by these new, "multiple-measure" evaluation systems aimed to be more differentiated, to provide better feedback to teachers about their practice, and to hold teachers accountable for their students' outcomes (Steinberg & Garrett, 2016). Presumably, they would also provide more actionable information for personnel decisions, such as removing ineffective teachers or compensating high performers, as well as for teacher development efforts (Donaldson & Papay, 2015).

Research on multiple-measure teacher evaluation reports mixed progress toward these goals. Despite implementation of rubrics that elaborate effective and ineffective teaching practice, classroom observation ratings identify relatively few teachers as demonstrating unsatisfactory teaching (Kraft & Gilmour, 2017), with principals showing reluctance to assign low ratings even to teachers they believe are ineffective (Grissom & Loeb, 2017). Nonetheless, principals report making use of evaluation data for teacher development

(T. A. Drake et al., 2015), and evidence suggests that principals are basing their efforts to retain some teachers and move others out of their school on evaluation information as well (Adnot, Dee, Katz, & Wyckoff, 2017; Dee & Wyckoff, 2015; S. Drake et al., 2019; Grissom & Bartanen, 2019).

The degree to which teacher evaluation systems can affect personnel decisions and teacher development in positive ways depends on the accuracy of the information those systems provide. Evaluation ratings that produce invalid signals of teacher performance can lead to unproductive personnel strategies and identification of incorrect targets for improvement. Bias is a threat to the accuracy of teacher effectiveness ratings produced by multiple-measure teacher evaluation systems. Of particular concern for our analysis is bias in teacher observation ratings, which make up roughly 50% of the overall evaluation score in the typical system (Steinberg & Garrett, 2016). All evaluation metrics are proxies that capture teacher performance with a degree of error; bias means systematic deviation of teacher observation scores from actual instructional effectiveness.

We are especially concerned with the question of whether teacher race/ethnicity and gender are associated with such systematic deviations. Next, we discuss possible sources of bias in teacher observation ratings and the ways in which teacher race/ethnicity and gender may be relevant for those potential sources.

Bias from Nonrandom Sorting of Teachers and Students

A relatively large body of research considers the impact of nonrandom sorting of teachers and students on teacher value-added measures (see Koedel et al., 2015, for a review). Bias can arise if students are assigned to teachers on the basis of factors not sufficiently accounted for in the value-added model. Teachers in schools whose neighborhoods have more violence or less community engagement, for example, may receive lower value-added scores because the students in that school are exposed to non-school factors that may negatively impact their achievement. Bias can also arise from

within-school sorting. For example, more motivated high school students may select course sections with teachers they think will challenge them more. If students' motivation predicts their test scores but are not captured well by the model, estimates of their teachers' value-added may be biased upward.

Bias from nonrandom sorting can extend to teacher observation (Cohen & Goldhaber, 2016). In this case, nonrandom sorting can lead to systematic differences between a teacher's "true" instructional performance and the score assigned in the observation. Assignment to some types of students may make it more challenging for a teacher to demonstrate satisfactory performance according to each of the various indicators enshrined in the observation rubric. For example, students who are several years below grade level in terms of their mastery of skills and course content may require that their teachers focus more on remediation and differentiated instruction. Similarly, students with a greater propensity to commit disciplinary infractions may require that their teachers focus more time and attention on classroom management. Importantly, the ability to deliver high-quality instruction to students with heterogeneous needs *is* an aspect of effective teaching, but may be less recognized or rewarded in high-stakes evaluations.[1] Teachers who work in schools with many more challenging students may thus receive lower evaluation scores. Within schools, if some teachers are systematically more likely to be assigned students with lower prior achievement or more discipline difficulties, their observation ratings similarly may be biased downward.

Observation rating bias from teacher and student sorting may affect teachers differently by gender and by race/ethnicity. Teachers from different subgroups sort differently across schools. Teachers of color, for example, teach in schools with higher fractions of low-income and low-achieving students, and students of color (Sun, 2018).

[1] By extension, we do not presuppose that teachers who work with struggling students or students with many disciplinary infractions cannot be highly effective or that a student's background makes them inherently "harder to teach." Instead, we argue that measures of teaching effectiveness may not be well-suited to accurately identify the performance of teachers who work with such students.

Within schools, teachers with different demographic backgrounds also may teach students with different characteristics. Teachers of color may be seen as disciplinarians, making it more likely that they are assigned students with disciplinary challenges (e.g., Brockenbrough, 2015). Black and Hispanic teachers are more likely to be assigned lower-achieving students of color (Kalogrides et al., 2013). In high schools, white teachers are more likely to be assigned to honors courses with larger numbers of high-ability students (Grissom, Kabourek, & Kramer, 2020). In their analysis of data from the MET project, in which students were randomly assigned to teachers within schools, Campbell and Ronfeldt (2018) found that teachers in classrooms with larger fractions of Black and Hispanic students received lower observation scores from trained raters, while higher average (incoming) student achievement was associated with higher ratings (see also Steinberg & Garrett, 2016). These findings suggest the potential for negative biases in Black and Hispanic teachers' observation ratings due to nonrandom sorting. Indeed, Campbell and Ronfeldt (2018) find that Black teachers' lower observation ratings in the MET data could be explained by classroom composition.

Predictions regarding teacher gender-related bias due to nonrandom student sorting are less clear. Female teachers are more likely to work in elementary schools, where analysis of the MET data found systematically higher observation ratings than those in upper grades (Mihaly & McCaffrey, 2014). On the other hand, within schools, female teachers may be more likely to be assigned lower-achieving students, especially special education students (Kalogrides et al., 2013), which may impact opportunities to score highly. Even with adjustments for classroom composition, Campbell and Ronfeldt (2018) find that female teachers receive higher ratings, on average.

Additional Potential Sources of Bias in Teacher Observation Scores

While bias from nonrandom sorting is the primary threat to interpreting teacher value-added measures as the average causal effect of a teacher on his or her students'

learning, observation scores are open to important potential sources of bias beyond sorting. Two of these are deficiencies in the rubrics or instruments used to rate teachers, and rater bias, that is, biases associated with the school leaders or other officials who complete the observation (Cohen & Goldhaber, 2016; Milanowski, 2017).

Rubrics employed in high-stakes teacher observations may be constructed in ways that lead to lower scores for teachers of students with some characteristics, or for some kinds of teachers. As Milanowski (2017) describes, common rubrics, such as the Framework for Teaching (FFT), may leave out behaviors that are effective with key student subgroups, such as low-achieving students, or assign those behaviors low scores. Students performing below grade level, for example, may require more direct or structured instruction to catch them up before they can benefit from teaching techniques that the rubric values more. This feature of the observation rubric may be one mechanism by which the nonrandom sorting of low-achieving students can bias teachers' observation scores. Relatedly, scholars have expressed concern that the general observation rubrics typically used for teacher evaluation are ill-suited to capturing effective instructional practices for special education students (Jones, 2016), which may result in downward bias of observation scores of teachers of special-needs students. Moreover, generic descriptors of good teaching practice enshrined in rubrics like the FFT may not describe instructional approaches that are equally effective in all subjects (e.g., Rink, 2013).

Observation rubrics may suffer from a more fundamental problem that leads to differentiation of scores for teachers from different demographic groups. Some scholars argue that the FFT and similar rubrics are based on ideas about effective teaching that devalue instructional approaches Black teachers are more likely to use, such as culturally relevant pedagogy (Salazar, 2018).

Rater biases, or the tendency of raters to incorporate nonperformance information into their ratings, may also be relevant. Despite the presumed objectivity that comes from well-elaborated standardized rubrics, rater subjectivity is inherent in teacher observation.

Observation is a complex, cognitively demanding process that requires raters to map the instruction they see onto multiple indicators whose descriptions may fit imperfectly, requiring them to make assumptions and draw on their experiences to apply the rubric (Cohen & Goldhaber, 2016). Raters' biases may inform this process, affecting teachers indirectly and directly. Indirectly, raters may hold biased views of some student subgroups that affect their interpretation of teachers' instructional or classroom management strategies in the presence of those students (Milanowski, 2017). More directly, raters may hold implicit or explicit biases against teachers with some background characteristics, perceiving male teachers or teachers of color as less effective, for example, which affects how they rate what they observe in those teachers' classrooms (Rinehart & Young, 1996). Such biases may be correlated with the rater's own race/ethnicity or gender, though need not be (Grissom & Loeb, 2017). Raters may also tend to give higher performance ratings to teachers with whom they have good interpersonal relationships or with whom they have worked together longer (Cohen & Goldhaber, 2016).

Data

This study analyzes administrative data from Tennessee, a state made up of 146 districts operating roughly 1,800 schools that serve 996,000 students. Data were made available through the Tennessee Education Research Alliance at Vanderbilt University with approval from the Tennessee Department of Education. Thirty-two percent of the state's students are Black or Hispanic, and 35% are economically disadvantaged.[2] Tennessee was a first-round winner of the Obama administration's Race to the Top competition and instituted a number of educational reforms under its auspices. These reforms included a requirement that all educators be evaluated via a multiple-measure evaluation system beginning in the 2011–12 school year. The state designed the Tennessee Educator Acceleration Model (TEAM) to meet this requirement, though districts could also use

[2] https://www.tn.gov/education/data/report-card.html

another system with state approval. The state approved three alternative systems (COACH, TEM, and TIGER), which have been used in a small number of districts. Because all four models have similar components, we focus on TEAM in our description, though in some analyses we look for differences across evaluation systems. In all systems, teachers' overall summative evaluation scores comprise the weighted average of three components: scores from formal classroom observations, student test score growth, and student achievement. Our analysis focuses on classroom observation scores, which receive the greatest weight in determining the summative rating.

Administrative data contain demographic, job classification, and location information for all K–12 public school employees. In each year we can access each educator's job title and placement, years of work experience in the state's school system, highest degree obtained (e.g., Master's degree, educational specialist), and salary. The data also include information on educator sex (binary, listed as female or male) and race/ethnicity (white, Black, Hispanic, Asian, Native American, or other). In Tennessee, the fraction of Asian, Native American, and other race/ethnicity educators was too small to permit a robust analysis, so teachers falling into these categories were dropped. Additionally, Tennessee's administrative files do not reliably identify Hispanic ethnicity in every year, forcing us to limit our analysis to Black and white teachers.[3] We merge the staff data with files containing teachers' evaluation information, which are available from 2011–12 through 2016–17. In addition to the average observation score that contributes to teachers' summative evaluation ratings, beginning in 2012–13 we can also access observation-level information (item scores, rater identifiers) for teachers in districts using the TEAM observation rubric, which are 82% of the state's teachers. Additionally, beginning in 2015–16, we can access observation-level information for teachers from Shelby County (one of the districts that uses an alternative observation rubric), which is important because it

[3] The 2011–12 Schools and Staffing Survey estimates that 97 percent of Tennessee teachers are Black or white.

has the largest number of Black teachers in the state.

Classroom observation scores are conducted by trained raters according to the rubric associated with their evaluation system (e.g., TEAM, COACH). The TEAM rubric, used by the vast majority of districts, defines levels of performance on 19 instructional indicators in the domains of instruction, environment, and planning, plus four additional indicators describing teacher professionalism.[4] Teachers receive scores of 1 ("significantly below expectations") to 5 ("significantly above expectations") on each indicator. The rubrics approved for the other systems cover different domains, though with substantial overlap with the content of the TEAM rubric. In Tennessee, teachers typically receive between two and five observations per year,[5] and more than 90% of observations are conducted by the school principal or assistant principal, with the remainder performed by central office officials or teacher observers. Scores averaged over the school year become the summative classroom observation rating.

Table 1 shows average teacher, school, colleague, and observation characteristics for Tennessee teachers. In addition to the means across all teachers, we also show these characteristics by race and gender. Eighty-eight percent of teachers are white and 79% are female. Similar to national trends, Tennessee's teacher workforce is far less racially diverse than the student population. While Black and white teachers in Tennessee have similar demographic characteristics, they work in very different school contexts. For instance, the average white teacher works in a school with 17% Black students, compared to 64% for the average Black teacher. Similarly, Black teachers systematically work in urban schools while white teachers are more evenly dispersed across locale types. Comparing male and female teachers, the main difference is school level. Roughly half of male teachers work in high schools, compared to only 19% of female teachers. Consistent with the patterns for student

[4] The TEAM rubrics are available at https://team-tn.org/evaluation/teacher-evaluation-2/

[5] State policy with respect to number of observations does not provide clear-cut requirements for the number of classroom visits a teacher must receive. Rather, policy sets minimum requirements for the number of times a teacher must be rated on a particular rubric domain (e.g., instruction), and often a single classroom observation yields scores on multiple domains.

demographics, the average white teacher works in a school with few Black teachers (7%) and is unlikely to have a Black principal (11%). While 61% of Black teachers work in a school with a Black principal, only 45% of their colleagues are Black, on average. Mainly due to the sorting by school level, men are more likely to have more male colleagues and are more likely to work for a male principal.

The bottom of Table 1 shows observation characteristics. Similar to other states, classroom observation scores are skewed, with most teachers falling between 3 and 5 on the 1 to 5 scale. The average score is 3.92 with a standard deviation of 0.59. To facilitate interpretation and ensure consistency across years, we standardize scores within each year. As mentioned above, not all teachers receive the same number of observations each year, though almost all receive between two and five. The average teacher is observed 3.1 times by 1.9 different raters, with no substantive differences by teacher race or gender.

Methods

We begin by documenting race and gender gaps in teachers' classroom observation scores by estimating the following model via OLS:

$$Score_{ist} = \beta_0 + \beta_1 BlackTch_i + \beta_2 MaleTch_i + \epsilon_{ist} \quad (1)$$

where the average observation score (standardized by year) of teacher i in school s in year t is regressed on indicators for male and Black. Negative coefficients for β_1 and β_2 indicate that Black and male teachers have lower average observation scores than white and female teachers, respectively. [6] In all models, we cluster standard errors at the school level.

The main focus of our analysis is to test various possible explanations for race and gender gaps in observation scores. Our general approach is to examine how β_1 and β_2

[6] Note that we can also include the interaction between Black and male. While we show the descriptive findings for this model, the bulk of our analysis focuses on race and gender gaps, rather than the intersection of race and gender.

change when we add different sets of covariates to equation 1:

$$Score_{ist} = \beta_0 + \beta_1 Black_i + \beta_2 Male_i + \delta X_{ist} + \mu_{st} + \epsilon_{ist} \quad (2)$$

where X may include teacher characteristics (age, education level, experience), school characteristics (student demographics, enrollment size, level, locale type), assigned student characteristics (demographics, prior-year test scores, attendance, and discipline), teaching assignment (grade level, subject), and individual value-added (according to the Tennessee Value-Added Assessment System, or TVAAS).

Although we have access to a rich set of school-level characteristics, there remains a concern that unobserved school factors could explain race and gender differences in observation scores, which we might otherwise attribute to alternative explanations, such as rater bias. To address this concern, we estimate models that include school-by-year fixed effects (μ_{st}), such that β_1 and β_2 are estimated only by comparing Black and white or male and female teachers who work in the same school in the same year. In addition to unobserved school-level heterogeneity, these models also account for any school-specific shocks, such as a principal transition or a change in the curriculum.

Beyond modeling observation scores at the teacher-by-year level, we can also leverage the fact that teachers have multiple observations over the course of the year. These observation-level data allow us, for instance, to examine the extent to which race and gender gaps vary by rubric domain or observation order. We can also examine whether race and gender gaps are explained by differences in rater characteristics. For TEAM teachers beginning in 2012–13 (plus those in Shelby County beginning in 2015–16), we estimate models of the following form:

$$Score_{ijnst} = \beta_0 + \beta_1 Black_i + \beta_2 Male_i + \delta T_{it} + \phi R_{jt} + \mu_{st} + \delta_j + \gamma_n + \epsilon_{ijnst} \quad (3)$$

where T are the teacher characteristics described above, R are characteristics of the rater

(race, gender, education level, experience, job title), μ_{st} are school-by-year fixed effects, and δ_j are rater fixed effects. School-by-year and rater fixed effects are identified given that 91% of schools have multiple raters in a given year. We also include indicators for observation order to account for trends in average ratings within each year.

Descriptive Gaps in Observation Scores

We begin our analysis by descriptively examining (i.e., without any adjustments for teacher, school, classroom, or rater characteristics) race and gender gaps in teacher-by-year average observation scores. Figure 1 shows these gaps for each year beginning in 2012—the first year that Tennessee implemented its multiple-measure teacher evaluation system. The top panels show average observation scores for race and gender separately, while the bottom panels show the four combinations of race and gender. We show both raw scores from the rubric (ranging from 1 to 5) and scores that are standardized by year. Several patterns are evident from the figure. First, in each year, white teachers receive higher average observation scores than Black teachers, and women receive higher average scores than men. Second, although average observation scores are increasing over time for all groups, race and gender gaps are fairly constant; gender gaps change almost none across years, and, despite some movement, the magnitude of the Black-white gap in 2017 is equal to 2014. The third pattern is that the male-female gap is larger than the Black-white gap. Pooling across all years, women outscore men by 0.32 SD, while white teachers outscore Black teachers by 0.15 SD. As a result, Black men are the lowest-scoring teachers, receiving scores approximately half a standard deviation lower than white women.[7]

Table 2 shows descriptive gaps in teacher-by-year observation scores across six categories of subgroups: school level, school locale, the teacher observation rubric used by

[7] The race and gender gaps are approximately additive, both descriptively and when tested via an interaction term in our regression models. We thus omit the interaction between race and gender in the models we present.

the district, the rubric domain, the teacher's primary subject taught[8], and years of experience. Panel A shows race and gender gaps and Panel B shows race-by-gender gaps. In both panels, the omitted group is white female. The patterns are strikingly consistent across all subgroups. Regardless of the school context, observation rubric used, rubric domain, subject taught, or experience level, Black teachers receive lower average scores than white teachers and male teachers receive lower average scores than female teachers. However, we do find significant heterogeneity in the magnitude of these gaps, particularly for race. For instance, the Black-white gap is almost twice as large in high schools (-0.21 SD) than in elementary schools (-0.11 SD). In terms of locale, the average Black teacher in an urban school scores only marginally lower (-0.02 SD) than the average white teacher, but Black teachers in town/rural schools score far lower than white teachers (-0.35 SD).

Race gaps in observation scores also vary in magnitude according to the district's observation rubric. The most commonly used TEAM rubric shows substantially larger race gaps than the other rubrics. Additionally, for those teachers, we can disaggregate scores by the four rubric domains: instruction, environment, planning, and professionalism. We find that gaps in observation scores exist and are similarly sized across all four domains.

For subject taught, the largest race gap is for social studies teachers, with relatively smaller gaps for health/P.E., math, and self-contained teachers. The Black-white gap is substantially larger among teachers with more than 20 years of experience (-0.28 SD) and also slightly larger among brand-new teachers (-0.18 SD for 0–1 years of experience).

Gender gaps are less variable in magnitude across subgroups, except for teacher experience. For instance, the male-female gap is -0.25 SD, -0.30 SD, and -0.29 SD in elementary, middle, and high schools, respectively. Similar to the Black-white gap, the male-female gap is largest in town/rural schools (-0.35 SD), though there is also a sizable gap in urban schools (-0.26 SD). Subject taught and teacher experience show the greatest variability in the magnitude of the gender gap. The gap is largest for math teachers (-0.42

[8] We include teachers in a particular subgroup if 50% or more of their assignment was in the given subject.

SD) and smallest for arts/music teachers (-0.11 SD). The gender gap also steadily grows across the experience distribution, from -0.19 SD among first- and second-year teachers to -0.42 SD among teachers with more than 20 years of experience.

Turning to race-by-gender in Panel B, we observe that, relative to white women, Black women tend to have the smallest gap, while Black men often score lower than white women by more than half of a standard deviation. The largest observation score gap is in town/rural schools, where Black men receive scores that are 0.74 SD lower than white women, on average.

What Explains Observation Score Gaps?

The previous section establishes that there are large differences in average observation scores along race and gender lines. The remainder of our analysis aims to identify factors that explain these gaps.

Differences in Teacher Characteristics and School Context

We begin by examining the extent to which differences in teacher characteristics and school context explain race and gender gaps in observation scores. The results are shown in Table 3. In each column, the focal coefficients are Black and male teacher, which represent the Black-white and male-female observations score gaps. Column 1 shows the baseline race (-0.15 SD) and gender gaps (-0.32 SD). Column 2 adds controls for teacher education, age, and experience. Perhaps unsurprisingly, observation scores and highest education level are positively associated. For example, teachers with a master's degree outscore teachers with a bachelor's degree by 0.12 SD, on average. Conditional on experience, older teachers tend to receive lower scores than younger teachers. Finally, we observe a fairly steep experience gradient; compared to teachers with 15–24 years of prior experience, novice teachers (0–4 years) score half of a standard deviation lower. Controlling for teacher characteristics, we find that the Black-white gap increases in magnitude to -0.18 SD, while the male-female gap decreases to -0.29. The increase in the Black-white gap is explained by

Black teachers having higher levels of education in Tennessee, on average (experience and age are very similar). On the other hand, Tennessee's male teachers are slightly less educated and less experienced, on average. On the whole, however, teacher characteristics explain little of observation score gaps by race and gender.

Next, we add controls for school characteristics, including enrollment size, student demographics, school level, and locale type. As mentioned above, the average Black and white teacher in Tennessee work in very different school contexts. For example, one of the largest differences is in the proportion of Black students in the school. The average Black teacher works in a school where 64% of students are Black, compared to only 17% for the average white teacher (see Table 1). Also, most Black teachers work in urban schools, while the majority of white teachers work in town or rural schools. Large differences in school context may matter to the extent that observation scores implicitly measure school-level factors. Prior studies, for instance, have shown that teachers' subjective evaluation scores in part capture the demographic characteristics of the students they teach (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). Controlling for school characteristics, then, helps uncover the extent to which race gaps are explained by teacher sorting patterns.

As shown in column 3, adding these controls reduces the Black-white gap from -0.18 to -0.04 SD. Examining the estimated coefficients for school characteristics confirms that there is a substantial relationship between school context and observation scores. On average, teachers receive lower observation scores in schools with more Black and Hispanic students, fewer gifted students, and more students qualifying for free/reduced-price lunch. Teachers in middle and high schools receive lower average scores than those in elementary schools. Conditional on student demographics and school level, there are no significant differences among teachers from different locale types. The gender gap also decreases slightly in magnitude when accounting for school characteristics (-0.29 SD to -0.26 SD). The decrease is explained by a single factor: men are much more likely to work in high schools (51% to 19%), where teachers receive systematically lower observation scores.

Taken at face value, the estimates in column 3 imply that comparing Black and white teachers who work in similar school contexts yields almost no gap in average observation scores. We might be tempted to conclude, then, that the primary driver of race gaps in observation scores is teacher sorting, rather than another sort of systematic bias against Black teachers, such as rater bias. However, when we shift to our preferred specification in column 4, which replaces school characteristics with school-by-year fixed effects, we recover a Black-white gap that is similar in magnitude to columns 1 and 2. The estimated male-female gap is essentially identical in columns 3 and 4.

What might explain the large difference in the Black teacher coefficients for these specifications? Our reanalysis by race and gender subgroups (shown in Appendix Table A1) uncovers that the model in column 3 is misspecified; specifically, there exists substantial heterogeneity in the relationships between school characteristics and observation scores for Black versus white teachers. When estimating separate models for Black and white teachers, the relationships between observation scores and school characteristics are very different. Most notably, the *Proportion Black Students* coefficient is 0.40 for Black teachers and -0.25 for white teachers. This large difference in the slope, combined with the fact that the average Black and white teacher work in schools with very different proportions of Black students, means that the bias from an omitted interaction (i.e., *Black Teacher* $\times$ *Proportion Black Students*) is substantial.[9]

Columns 5 and 6 in Table 3 add an interaction between Black teacher and the proportion of Black students in the school. Regardless of whether we control for school characteristics (column 5) or school-by-year fixed effects (column 6), the interaction term is positive, large in magnitude, and statistically significant. In substantive terms, the *Black Teacher* $\times$ *Prop. Black Students* coefficient in column 6 demonstrates that the Black-white

[9] While there are also substantive differences in the coefficients for some of the other school characteristics (e.g., proportion of Hispanic students, proportion of gifted students), the magnitude of bias from omitted interactions is much smaller because the correlation between teacher race and these other characteristics is much smaller.

gap in observation scores is largest in schools that have few Black students and smallest in schools with many Black students.

Exploring Heterogeneity in the Black-White Observation Score Gap

In Table 4, we further examine student race as a moderator of the Black-white gap in teacher observation scores. Specifically, we test whether student racial composition is a proxy for other factors, such as the racial composition of the teaching staff or school administration. Columns 1, 2, and 3 estimate interactions between Black teacher and the school's proportion of Black students, colleagues (i.e., other teachers in the school), and administrators (combining principals and assistant principals), respectively. When included separately, each interaction is statistically significant, in the expected direction, and large in magnitude. Column 2, for instance, shows that the estimated Black-white gap decreases by 0.39 SD moving from a school with a single Black teacher to a school with all Black teachers. When we include student, colleague, and administrator demographics in the same model (column 4), we find a striking result: the large, positive interaction between Black teacher and proportion of Black students is completely attenuated, while the interactions for Black colleagues and administrators remain positive and statistically significant. In other words, the shrinking Black-white gap in schools with more Black students appears to be explained by the fact that there are more Black colleagues and administrators in those schools. In particular, colleague race remains a salient moderator of the Black-white gap.

To further illuminate the dynamics in Table 4, Figure 2 plots the estimated Black-white gap in observation scores as a function of the proportion of Black colleagues in the school. We show estimates from four different specifications, all of which include teacher controls and the interaction between Black teacher and the proportion of Black colleagues in the school. Importantly, we estimate the relationship non-parametrically (instead of assuming a linear relationship) by dividing the proportion of Black colleagues into categories. Panel A controls only for the proportion of Black colleagues and includes

no other school characteristics, while Panel B control for school characteristics. Panels C and D replace school characteristics with school-by-year FE, and Panel D also includes interactions between Black teacher and the proportion of Black students and Black administrators, respectively. Across all specifications, we find a consistent pattern: the Black-white gap in observation scores narrows in schools that have more Black teachers. In our preferred specification that includes school-by-year fixed effects (Panel C), for example, the Black-white gap ranges from roughly -0.25 SD in schools with 0–10% Black colleagues to zero in schools with a majority of Black colleagues.

Figure 3 shows the distribution of colleague race (i.e., what proportion of a teacher's colleagues are Black) for Black and white teachers in Tennessee. The left plot shows the distribution for the full sample, and the right plot shows the distribution for the effective sample, in the school-by-year FE model—defined as school-by-year cells where there is at least one Black teacher. The vast majority of white teachers work in schools with few or no Black colleagues, while relatively even proportions of Black teachers work in schools that are racially isolated or mixed. Based on panel C in Figure 2, the mean Black teacher works in a school with 45% Black colleagues and a predicted Black-white gap of -0.10 SD, while the mean white teacher in the effective sample works in a school with 13% Black colleagues and a predicted Black-white gap of -0.25 SD. That said, roughly half of Black teachers in Tennessee work in a school where the predicted Black-white gap is zero.

One reason why having more Black colleagues might matter for the Black-white gap is that raters likely make implicit comparisons among teachers, and being the only Black teacher in a school or grade level makes it more likely that he or she is compared primarily to white teachers, whose practices may be taken as the benchmark for "high-quality" teaching in that school. If, for instance, Black teachers are more likely to implement alternative pedagogical or classroom management approaches (e.g., culturally relevant pedagogy, restorative justice), such practices may not be deemed as "effective" if other teachers do not use them. In columns 5 and 6, we further examine this relationship by

disaggregating colleague race into two groups: teachers in the same grade as the focal teacher and teachers in different grades. If raters are making explicit comparisons among teachers who instruct the same groups of students, we would expect that the interaction between teacher race and colleague race is driven by same-grade colleagues rather than other-grade colleagues. However, columns 5 and 6 show that the racial composition of both same-grade and other-grade colleagues matters, as both of the interactions are positive and statistically significant.

The results in Table 4 inform our approach to investigating the Black-white observation gap throughout the rest of the paper. Specifically, given our interest in evaluating potential explanations for these gaps, it is important to estimate a model that accounts for heterogeneity in the race gap. Therefore, our subsequent analyses model this heterogeneity by including the interaction between Black teacher and the proportion of Black colleagues in the school.[10] Including this interaction term changes the interpretation of the *Black Teacher* coefficient. Instead of capturing the predicted difference between Black and white teachers, on average, it represents the predicted difference between Black and white teachers in a school with no Black colleagues. To provide a more meaningful sense of the magnitude of the Black-white gap while maintaining parsimony, we instead report the estimated marginal effect in a school with 20% Black colleagues, which is the mean proportion among teachers in the effective sample (i.e., teachers in school-by-year cells that have at least one Black teacher).[11]

---

[10] Given the high correlations between proportion of Black students, colleagues, and administrators, our findings are very similar if we instead include an interaction with Black students or Black administrators, or include all three interactions. For the sake of parsimony and precision, we only model the interaction between Black teacher and proportion of Black colleagues in the school.

[11] An alternative is to report both the main effect (*Black Teacher*) and interaction term (*Black Teacher x Proportion of Black Colleagues*) for each specification. However, this approach adds complexity, and we found that it yields little additional insight relative to simply reporting the marginal effect.

Differences in Teacher Value-Added

Another potential explanation of the race and gender gaps is that they reflect true differences in teacher effectiveness. In other words, white and female teachers may receive higher observation scores because they are more effective than Black and male teachers, on average. Relatedly, heterogeneity in the Black-white gap may reflect the tendency for higher-quality teachers to sort into schools with more same-race colleagues or administrators. While we cannot observe true instructional quality, we can shed light on these possibilities by leveraging alternative measures of teacher effectiveness. Specifically, we can use estimates of teachers' impacts on student test scores, or value-added (VA). As part of the teacher evaluation system, Tennessee contracts with the SAS Institute to produce VA estimates for individual teachers who are in tested grades and subjects. This system is called the Tennessee Value-Added Assessment System (TVAAS). In addition to accessing TVAAS scores, we can also estimate our own teacher VA models using the same test score and linkage files provided to SAS to estimate TVAAS. Specifically, we follow the leave-year-out, drift-adjusted approach outlined in Chetty et al. (2014).[12]

In Table 5, we examine the extent to which race and gender gaps in observation scores are explained by differences in VA. We estimate models with and without school-by-year FE, and we restrict the sample to teachers for whom we can estimate VA. Columns 1 and 5 show the baseline gaps, with column 5 including school-by-year FE and the interaction between Black teacher and proportion of Black teachers in the school. In models that do not adjust for differences in school context, the change in the estimated Black-white observation score gap depends on which VA measure we include. When

[12] The estimation steps are as follows. First, we residualize student test scores (separately by subject) on a vector of prior-year test scores, student characteristics (race/ethnicity, gender, FRPL eligibility, gifted status, special education status, lagged absences, grade repetition, and whether the student changed schools at least once during the year), school- and grade-level averages of these student characteristics, grade-by-year fixed effects, and teacher fixed effects. After computing the student residuals, we add back the teacher fixed effects and estimate the best linear predictor of a teacher's average student residuals in the current year based on their residuals from prior and future years. The coefficients from this best linear predictor are then used to predict a teacher's value-added in the current year.

controlling for drift-adjusted VA in column 2, the Black-white increases in magnitude, whereas it shrinks in column 3 when controlling for TVAAS. Regardless of which VA measure we include, the male-female gap shrinks by roughly 15%.

In our preferred specification (school-by-year FE) in columns 5–8, we find more consistent results between models that include drift-adjusted VA versus TVAAS. For both the Black-white and male-female gaps, adjusting for teacher VA only slightly reduces the gap relative to the baseline model. In other words, we find that only a small portion of observation score gaps by race or gender can be explained by differences in teachers' contributions to student achievement. [13]

Differences in Student Assignment

Next, we examine the extent to which within-school observation score gaps are explained by differences in the composition of students assigned to teachers. Prior work has demonstrated that teachers who are assigned higher proportions of Black, Hispanic/Latino, and low-achieving students tend to receive lower observation ratings (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016). Importantly, Campbell and Ronfeldt (2018) document that this pattern holds even when students are randomly assigned to teachers within a school, which in their study rules out the possibility that these observation score gaps reflect true differences in teacher quality. We first document gaps in student assignment by teacher race and gender, then examine whether accounting for these differences in student composition explain within-school gaps in observation scores.

[13] There are at least three reasons why drift-adjusted VA and TVAAS are independently predictive of observation scores. First, TVAAS incorporates student performance from the current year, while drift-adjusted VA does not by construction. To the extent that idiosyncratic variation in classroom performance is captured by both TVAAS and observation scores (e.g., having an unusually motivated group of students), TVAAS will be correlated with observation scores even conditional on drift-adjusted VA. Second, there are differences in how TVAAS and drift-adjusted VA account for student sorting. For instance, TVAAS does not control for students' demographic characteristics, while drift-adjusted VA does. If observation scores are correlated with student demographics, then TVAAS and drift-adjusted VA will be independently predictive of observation scores. Finally, drift-adjusted VA incorporates more years of test score data than TVAAS, including future years. To the extent that past and future performance helps to predict teachers' current year performance, drift-adjusted VA will be predictive of observation scores even conditional on TVAAS.

Table 6 shows race and gender gaps in student assignments. In each regression, the dependent variable is listed in the column header. For student demographics, gaps represent differences in the average proportion of students assigned to a teacher. The in-school suspension (ISS) and out-of-school suspension (OSS) columns show differences in the proportion of teachers' students who had one or more of these suspensions in the prior year. Absences, math, and ELA are standardized by grade and year at the student level, then averaged across teacher assignments. We also provide the means of within-school standard deviations of each dependent variable. Because there tends to be far less variation in assigned student characteristics within a given school than across the entire state, this within-school SD provides a sense of the magnitude of race and gender gaps in student assignment relative to the typical amount of within-school student sorting. We include school-by-year FE to restrict the comparisons to teachers who work in the same school in the same year. Additionally, we include the interaction between Black teacher and the proportion of Black colleagues in the school.

The first five columns show within-school differences in students' demographic characteristics. We find that Black teachers are systematically assigned more historically disadvantaged students, particularly in schools where they have few Black colleagues. Relative to their white colleagues, Black teachers are assigned more Black students, more FRPL-eligible students, fewer gifted students, and more students receiving special education services. For instance, in a school with 20% Black colleagues, a given Black teacher is assigned 2.5 percentage points more Black students than their white colleagues, on average. While this difference is modest in absolute terms, it is almost half of the within-school standard deviation, meaning that the Black-white difference in assigned Black students is large relative to the typical amount of student sorting in a school. Turning to the interaction term, we also see that these gaps in student demographics decrease in schools with more Black students. This pattern suggests that within-school assignment differences could partially explain why Black-white observation score gaps are

smaller in schools with more Black students.

Columns 6–10 show gaps based on assigned students' prior-year outcomes. We find essentially the same pattern; Black teachers have greater proportions of students who were suspended in the prior year (both in-school and out-of-school suspensions), students who were absent more often, and students with lower prior-year test scores in math and ELA. Interestingly, the assignment gap between Black and white teachers for out-of-school suspensions actually increases in schools with more Black students, though the gaps in prior test scores decrease. One potential explanation for this pattern is that principals internalize that Black students (who have higher rates of suspension) have better behavioral outcomes when assigned to a Black teacher (Lindsay & Hart, 2017).

We also find some differences in student assignment between men and women. The largest gap is in the proportion of female students; men teach 4.7 percentage points fewer female students, on average, which is nearly half of the within-school standard deviation. Differences in other student demographics are smaller, but men tend to teach fewer disadvantaged students. Men are also assigned slightly higher-performing students in math and ELA but teach greater proportions of students with prior-year suspensions, suggesting that while they receive favorable assignments overall, principals may perceive men as better disciplinarians, which leads them to assign more students with behavioral issues.[14]

Table 7 examines the extent to which the differences in within-school student assignment shown in Table 6 explain observation score gaps. In each model, we estimate our preferred specification that includes teacher characteristics, school-by-year FE, and an interaction between the indicator for Black teacher and the proportion of Black colleagues in the school. The estimated Black-white gap is equal to the marginal effect of Black

[14] One question raised by the patterns in Table 6 is how much of these assignment gaps are driven by a single demographic difference, such as the student gender gap for male teachers and the student race gap for Black teachers. As shown in Table A2, Black-white gaps in student assignment still appear when controlling for the proportion of Black students, though they are attenuated. Table A3 shows that after controlling for the proportion of female students, male teachers have even larger advantages in terms of student assignment.

teacher in a school with 20% Black colleagues. The estimated male-female gap is simply the coefficient for male teachers from the regression model. Column 1 shows the baseline gaps in observation scores for the sample of teachers that have student assignment information (not including prior-year test scores). Column 2 introduces controls for assigned student characteristics. Even within the same school and year, the characteristics of students taught predict observation scores. Teachers who are assigned more traditionally disadvantaged students tend to receive lower observation scores. Additionally, after controlling for assigned student characteristics, the Black-white gap decreases from -0.19 to -0.15 SD, while the male-female gap decreases marginally, from -0.25 to -0.24 SD.

Columns 3–5 examine whether including additional controls for the prior-year scores of assigned students further narrows the race and gender gaps in observation scores. Comparing columns 4 and 5, we find that after controlling for demographics and prior-year suspensions and absences, adding prior-year test scores only slightly increases model fit and does not appreciably change the Black-white or male-female gap. Overall, the results in Tables 6 and 7 demonstrate that there are average differences in the characteristics of students assigned to Black and male teachers compared to white and female teachers within the same school, and that these assignment differences do explain some of the gap in observation scores, particularly by teacher race. However, even accounting for student assignments, substantial race and gender gaps remain.

Differences in Subject and Grade Assignment

Table 8 examines the extent to which subject and grade assignments explain observation score gaps. Column 1 is the baseline—it is equivalent to the specification shown in Table 7 column 2. In column 2, we add to the model controls for the subjects taught by each teacher. Because teachers can have multiple subject assignments, we operationalize subject taught in proportional terms, with full (100%) ELA teachers as the reference category. We do find that subject taught is correlated with observation scores.

For instance, while math teachers tend to score slightly higher than ELA teachers, science, social studies, and self-contained teachers score substantially lower. Among non-core subjects, foreign language and health/P.E. have lower average scores than ELA teachers, with no significant differences for career/technical education or arts/music teachers. However, controlling for subject taught has no appreciable effect on the size of the Black-white or male-female gap in average observation scores. Column 3 includes controls for the grade level of teachers' assigned students. Again, we operationalize these variables as proportions, and the reference category is ninth grade. Similar to subject, grade assignments are predictive of observation scores but do not explain race or gender gaps. When controlling for both subject and grade taught in column 4, the race and gender gaps decrease by only 0.002 SD.

Differences in Rater Characteristics

Next, we consider the extent to which rater characteristics may influence race and gender gaps in teacher observation scores. Here, we leverage the fact that teachers receive multiple classroom observations each year, which often are conducted by different raters. School principals and assistant principals conduct over 90% of observations, with the remainder split between central office personnel and teacher observers. The average teacher has two different raters in a given year. In addition to current job title, we can observe raters' demographic characteristics, education level, and job history. Table 9, column 1 shows the baseline race and gender gaps from our preferred specification, which includes school-by-year FE and controls for teacher characteristics, assigned student characteristics, and subject/grade assignment. Whereas the unit of observation in prior tables is teacher-by-year, we now shift to teacher-by-year-by-observation. Adding rater characteristics in column 2, there is no change in the estimated gaps, though some of the rater characteristics are predictive of observation scores. For instance, central office raters give substantially lower scores than principals, APs, or teachers, and raters with 10 or more

years as an administrator give slightly higher scores than those with fewer than three years, on average.

Two potential confounding factors are the order of observations and the total number of observations a teachers receives during the school year. As mentioned above, teachers in Tennessee typically receive between two and five observations in a given year, which is determined by a combination of prior-year evaluation rating, certification status, and district policy. As shown in column 3, both order and total number of observations are strongly correlated with a teacher observation scores. Specifically, teachers tend to receive higher scores with each subsequent observation. The average increase is 0.15 SD from the first to the second observation, and up to 0.59 SD for the fifth or higher observation. However, teachers who receive more observations during the year have substantially lower average scores than teachers with fewer observations. Compared to a teacher observed twice in a year[15], a teacher receiving five or more observations scores a full standard deviation lower, on average. This fact is unsurprising given that total number of observations correlates with teacher experience and prior effectiveness. Including these characteristics slightly lowers both the Black-white (-0.13 to -0.11 SD) and male-female (-0.21 to -0.19 SD) gaps. We also observe some changes in the estimated coefficients for the rater characteristics. For example, while principals gave the highest average scores according to column 2, both assistant principals and teacher raters give higher scores after adjusting for ordering and total number of observations. Additionally, the negative association between central office rater and observation scores has decreased in magnitude by more than half. Together, these results suggest that principals are more likely to observe teachers later in the year and/or teachers with fewer total observations, which explains why they give the highest predicted scores in column 2.

In column 4 we add rater fixed effects. If unobserved characteristics of raters are

[15] Less than one percent of teachers have only a single observation in a year, so we group one and two observations together for simplicity. All of our results are robust to dropping these teachers or including a separate indicator in the model.

contributing to race or gender gaps, including rater fixed effects will account for them to extent that they are fixed over time. For instance, this approach would account for a scenario where gaps are driven by Black or male teachers being systematically observed by harsher raters (i.e., raters that give lower average ratings regardless of teacher race or gender). However, we find little change in the race or gender gaps between columns 3 and 4.

In column 5, we look for evidence of teacher-rater matching effects. We find no benefit of having a same-gender rater—the estimated coefficient is a precise zero. However, we do find evidence of an effect for race: observation scores are 0.03 SD higher when the teacher and rater are the same race. It is important to note that with only two racial/ethnic groups, we cannot identify separate matching effects for Black and white teachers. Given that this race-match effect is relatively small in magnitude, it is perhaps unsurprising that its inclusion only reduces the estimated Black-white gap by a small amount (0.01 SD). We might expect, however, that the magnitude of the race-match effect varies by the racial composition of the school. Appendix Table A4 shows the results of re-estimating column 5 for subsamples of teachers in schools with 0–10%, 10–30%, and 30–100% Black colleagues, respectively. We find little evidence that the teacher-rater race match effect varies across these subsamples.

Even absent heterogeneity in the race-match effect, the tendency for Black (white) teachers to work in schools with Black (white) raters means that adjusting for teacher-rater race matching should differentially affect the size/direction of the Black-white gap as a function of the share of Black teachers in the school. Put another way, race-matching should favor white teachers (on average) in majority-white schools and favor Black teachers in majority-Black schools. We show this dynamic in Appendix Figure A1. Relative to the baseline model, adjusting for teacher-rater race matching reduces the size of the Black-white gap in schools with few Black teachers, since white teachers are substantially more likely to have a same-race rater in these schools. As the proportion of Black teachers in the school increases, the pattern flips—the race match effect serves to increase the

average scores of Black teachers relative to their white colleagues.

Teacher-Principal Work History

In Table 10, we examine whether race and gender gaps vary by characteristics of the work history between a teacher and principal. In column 1, we show the baseline gaps for observations conducted only by the teacher's principal.[16] In column 2, we test whether bias against Black teachers is lower among teachers who were hired by the principal conducting the observation. To the extent that rater bias exists, we might expect that it would be lower among principals who chose to hire that teacher. Column 2 supports this hypothesis; the average Black-white gap among teachers inherited by the current principal is -0.16 SD, compared to only -0.10 SD (-0.156 + 0.052) among teachers hired by the current principal. We find little evidence of this phenomenon for gender in column 4.

Columns 3 and 5 examine whether race and gender gaps change as a function of how long the teacher and principal have worked together. Although observation scores tend to be higher among teachers who have worked longer with their current principal, we find no consistent evidence that this relationship varies by teacher race.

Do Gaps Increase Over the School Year?

In Table 11 we examine whether race and gender gaps increase or decrease over the course of the year. Column 1 shows that, on average, observation scores tend to increase over the year. For example, average scores increase by 0.16 SD from the first to the second observation, up to almost 0.60 SD for the fifth or later observation. However, column 2 shows that Black teachers receive less of a bump; the Black-white gap is smallest for the first observation of the year at only 0.08 SD and increases by roughly the same amount for all subsequent observations.

[16] Note that we since we have restricted the sample to observations conducted by principals, we cannot include rater FE in these models, as they are perfectly collinear with the school-by-year FE.

In column 3, we distinguish between overall observation order and the number of times the same rater has conducted the observation. We find that some of the average increase in scores with additional observations is explained by the tendency for specific raters to give higher scores when they have observed a teacher more times in a year. In other words, while teachers receive higher average scores with each additional observation, they gain even more when these observations are conducted by the same rater. For example, the marginal return to a second observation with the same rater is 0.09 SD, up to 0.24 SD in the fifth or later observation. Importantly, we see that Black teachers benefit even more from being observed by the same rater. Comparing the magnitudes of the interactions for observation order (overall) and within-rater observation order, we find that the widening of the Black-white gap in later observations is greatly mitigated when Black teachers have the same rater throughout the year.

On the other hand, we do not find consistent evidence that gender gaps change over the course of the year, or that the returns to additional observations with the same rater are different between men and women. While a few of the estimated interactions in column 5 are statistically significant, they are small in magnitude and show no clear pattern.

Teacher Attitudes Towards Evaluation System

The previous sections demonstrate that there are substantial gaps in observation scores along racial and gender lines. As a supplementary analysis, we explore teachers' perceptions of the evaluation system in Tennessee, which come from a yearly statewide survey of educators called the Tennessee Educator Survey. We examine three of these survey outcomes in Table 12, which ask for teachers' perceptions of the fairness, usefulness, and onerousness of the evaluation system. These responses are available for the 2014–15 through 2016–17 school years.[17] Each measure has been standardized by year. We

[17] For each measure, we construct a standardized score. Panel A comes from a four-point Likert scale response ("strongly disagree", "disagree", "agree", "strongly agree"). In Panel B, the outcome is the factor score from teachers' responses on two four-point Likert scale items: "In general, the teacher evaluation process used in my school has led to improvements in my teaching." and "In general, the teacher evaluation

hypothesize that, given the large race and gender gaps in scores that remain unexplained by many observable factors, Black and male teachers will perceive the evaluation system less positively than their colleagues.

This hypothesis is not borne out. Panel A shows the results for fairness. In column 1, which only includes controls for teacher characteristics, there are no significant differences between Black and white or male and female teachers. Column 2 adds school-by-year fixed effects. Here we find that Black teachers rate the evaluation system as fairer than do their white colleagues in the same school, despite the fact that they receive systematically lower observation scores. Alternatively, we might expect that Black and male teachers perceive the evaluation process as less fair if they are in a school where the race or gender gap is larger. To examine this possibility, we first estimate school-by-year-specific coefficients for Black and male teacher, respectively, then include these coefficients as interactions in the perceptions models.[18] Column 3 shows that Black and male teachers' perceptions of fairness indeed are related to the size of the race and gender gaps in their schools. The positive interactions demonstrate that as Black and male teachers score closer to their white and female colleagues, they perceive the evaluation process as relatively fairer. However, the positive main effects of *Black Teacher* and *Male Teacher* indicate that even in schools where the race and gender gaps are zero, Black and male teachers have more positive perceptions than white and female teachers, respectively. Column 4 controls for the teacher's average observation score. If the interactions in column 3 truly reflect responsiveness to bias in observation scores, these coefficients should be attenuated when we control for teachers' scores, which is precisely what we observe. Unsurprisingly, teachers who receive higher scores perceive the evaluation system as fairer.

---

process used in my school has led to improvements in student learning." In Panel C, the question asks teachers to rate the following statement on a scale of 1 (high) to 5 (low): "The evaluation system is a burden."

[18] Our model to estimate school-by-year-specific slopes for Black and male teacher includes teacher characteristics and school-by-year fixed effects, which is the specification shown in Table 3 column 4. When including these estimates in columns 3 and 4 in Table 12, we report school-level cluster bootstrapped standard errors.

Panels B and C repeat this exercise for teachers' perceptions of the helpfulness of the evaluation system for improving their teaching and the degree to which the evaluation system is a burden, respectively. We generally observe the same patterns. Black teachers perceive the evaluation process as more helpful and less burdensome than their white colleagues, on average, but their perceptions are less favorable in schools where apparent racial bias is larger. Similarly, men rate the evaluation process as more helpful and less burdensome than women. Overall, these results suggest that while Black and male teachers tend to provide more positive assessments of the evaluation process, on average, they do appear to internalize apparent bias in their observation scores.

Discussion and Conclusions

As in prior research in low-stakes settings (Campbell & Ronfeldt, 2018), our analysis of classroom observations conducted as part of Tennessee's statewide teacher evaluation system finds large differences in the observation ratings assigned to teachers according to their race and gender. Black teachers score 0.15 SD lower than their white colleagues, and men score 0.32 SD lower than women, on average.

Our investigations of these two gaps yield different results, which are summarized in Table 13.[19] Our findings regarding the Black-white evaluation gap are nuanced. First, the average gap masks substantial heterogeneity by school context; Black teachers score substantially lower than white teachers in schools where they are racially isolated. As the percentage of Black colleagues increases, the Black-white gap narrows. Second, sorting of

[19] Specifically, we estimate models on a common sample of teachers with non-missing covariates to examine how much of the descriptive gap remains when we include the full set of controls. The sample (at the teacher-by-year-by-observation level) includes teachers for whom we can estimate value-added, with results for a broader set of teachers in Appendix Table A5. The sample also conditions on having observation-level data and non-missing information for teacher characteristics, assigned student characteristics, subject/grade assignment, and rater characteristics. All models include controls for observation order and total number of observations. Beginning in column 2, we add school-by-year fixed effects and then individually add sets of controls, with a fully saturated model in column 8. For the Black-white gap, we directly model heterogeneity to estimate the gap in schools with 0–25%, 25–50%, and 50–100% Black colleagues, respectively. These categories, respectively, include 31%, 21%, and 48% (92%, 5%, 3%) of Black (white) teachers in the state.

teachers across schools partially explains the Black-white gap, though substantial differences remain even when we limit to comparisons of Black and white teachers working in the same school in the same year. Third, although we find striking evidence that, within schools, Black teachers are assigned more students of color, low-income students, special education students, and students with histories of lower achievement, lower attendance, and greater disciplinary action, we do not find that accounting for differences in classroom composition explains the Black-white evaluation gap in the average school. This finding marks a departure from the conclusions of Campbell and Ronfeldt (2018), whose analysis of the MET data finds that differences in observation ratings among Black and white teachers become statistically indistinguishable from zero once they control for student characteristics.

We also find that while teacher characteristics, subject/grade assignments, and teacher quality (as measured by test score value-added) are associated with classroom observation scores, they do not explain within-school racial differences in scores. Finally, leveraging variation within school and school year in the characteristics of raters, we find that Black teachers receive lower ratings when observed by a white rater, which explains a small portion of Black-white gap. In sum, comparing the empty (column 1) and fully saturated (column 8) models in Table 13, we can explain roughly half of the Black-white gap in classroom observation scores.

While the average gender gap in observation scores is larger than the race gap, Table 13 shows that less of the gender gap can be explained by observable factors. Again comparing column 1 and column 8, controlling for all observable factors reduces the male-female gap by roughly 24%. Put simply, we are relatively unsuccessful at explaining why male teachers score substantially lower than their female colleagues. We also find no evidence that teachers benefit from being observed by a rater of the same gender. Future work should continue to investigate the gender gap, perhaps using data that can uncover processes that are unobservable in our administrative data.

These findings have several implications. Perhaps most importantly, our results add to a small, but growing body of literature demonstrating that subjective evaluations of teacher performance in part measure factors outside of their control, including teacher characteristics, school context, and the characteristics of a teacher's assigned students. In high-stakes contexts like the one in this study, ratings can drive personnel decisions such as contract renewal and compensation. Addressing these sources of bias is important to ensure fair treatment of teachers across school contexts, teaching assignments, and teacher background characteristics.

How can the apparent biases we document be addressed? Some prior work has suggested that, to offset biases against teachers of some student subgroups, observation scores should be adjusted for classroom composition using regression, similar to the way that value-added scores are adjusted (Whitehurst et al., 2014). A drawback of this approach is that such regression-based adjustments could mask real differences in the instructional quality of teachers assigned to different kinds of classrooms (Cohen & Goldhaber, 2016). Although this approach could be explored further, our results suggest that such adjustments would not be enough to account for the negative bias in the observation scores of Black and male teachers. Gaps between these teachers and their white and female counterparts persist in our data even after accounting for school sorting and the characteristics of the students they teach.

Instead, our results suggest that gaps are driven by factors other than nonrandom sorting, which may include rater bias and bias in the rubric itself. Although we cannot differentiate these two sources, we do uncover suggestive evidence of rater bias, namely that observers rate teachers they themselves hired more favorably, and also give higher ratings to teachers of the same race. If rater biases are the primary source of the residual difference between Black and white teachers, for example, it may be necessary to address implicit racial biases among raters via training, or to provide better training on application of the observation rubric more generally so that raters' discretion factors less into the

scoring process. To this point, the gaps in ratings by teacher characteristics documented in the MET project, which employed raters with extensive training, were substantially smaller than those we show for Tennessee. If biases arise from the rubric the state employs, which could happen if the rubric assigns higher value to teaching practices associated with white or female teachers even when other practices are similarly effective, policymakers may need to consider adjustments to the rubric to ensure that it captures a broader range of high-quality practices.

Our study is limited in at least two ways. First, our data come from a single state evaluation system with its particular approach to implementation, including the rubrics it employs, how it trains raters, and the regulations and expectations it sets for how observations are conducted and how scores are used. External validity of our results would be reinforced by future studies of observation ratings in other contexts. Second, unlike studies from the MET project (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016), we cannot leverage randomization of students to teachers, which leaves our study more open to concerns that ratings gaps are driven by differences in actual teaching effectiveness in classrooms with some groups of students or among teachers with different characteristics. Future work making use of exogenous variation in student assignment may arrive at different estimates of the biases we explore, though the general consistency of our descriptive findings with those Campbell and Ronfeldt (2018) show suggest that only partial attenuation of our estimates could be expected.

We conclude by pointing out that while we have probed some of the drivers of gaps in classroom observation scores, much remains unexplained, particularly for teacher gender. Additional analyses aimed at a more complete understanding of Black-white and especially male-female differences in teacher observation ratings would be fruitful avenues for future research.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher Turnover, Teacher Quality, and Student Achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 1–23.

Albert Shanker Institute. (2015). The State of Teacher Diversity in American Education. , 126.

Brockenbrough, E. (2015). âĂIJThe Discipline StopâĂİ. *Education and Urban Society*, *47*(5), 499–522.

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

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, *104*(9), 2633–2679.

Cohen, J., & Goldhaber, D. (2016). Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations. *Educational Researcher*, *45*(6), 378–387.

Dee, T. S., & Wyckoff, J. H. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Donaldson, M. L., & Papay, J. P. (2015). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy* (pp. 174–193). New York, NY: Routledge.

Drake, S., Auletto, A., & Cowen, J. M. (2019). Grading Teachers: Race and Gender Differences in Low Evaluation Ratings and Teacher Employment Outcomes. *American Educational Research Journal*, *56*(5), 1800–1833.

Drake, T. A., Goldring, E., Grissom, J. A., Cannata, M., Neumerski, C. M., Rubin, M., & Schuermann, P. (2015). Development or dismissal? Exploring principals' use of teacher effectiveness data. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 116–130). Teachers College Press.

Grissom, J. A., & Bartanen, B. (2019). Strategic Retention: Principal Effectiveness and Teacher Turnover in Multiple-Measure Teacher Evaluation Systems. *American Educational Research Journal*, *56*(2), 514–555.

Grissom, J. A., Kabourek, S. E., & Kramer, J. W. (2020). Exposure to Same-Race/Ethnicity Teachers and Advanced Math Course-Taking in High School: Evidence from a Diverse Urban District. *Teachers College Record*.

Grissom, J. A., & Loeb, S. (2017). Assessing Principals' Assessments: Subjective Evaluations of Teacher Effectiveness in Low- and High-Stakes Environments. *Education Finance and Policy*, 1–49.

Grissom, J. A., & Youngs, P. (2016). *Improving teacher evaluation systems: Making the most of multiple measures*. New York, NY: Teachers College Press.

Jones, N. (2016). Special education teacher evaluation: An examination of critical issues and recommendations for practice. In J. A. Grissom & P. Youngs (Eds.), *Improving teacher evaluation systems: Making the most of multiple measures* (pp. 116–130).

New York, NY: Teachers College Press.
Kalogrides, D., Loeb, S., & Béteille, T. (2013). Systematic Sorting: Teacher Characteristics and Class Assignments. *Sociology of Education*, *86*(2), 103–123.
Kane, T. J., & Staiger, D. O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation* (Tech. Rep.). Cambridge, MA: National Bureau of Economic Research.
Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195.
Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the Widget Effect: Teacher Evaluation Reforms ans the Distribution of Teacher Effectiveness. *Educational Researcher*, *46*(5), 234–249.
Lindsay, C. A., & Hart, C. M. D. (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. (2014). Grade level variation in observational measures of teacher effectiveness. In K. Kerr, R. C. Pianta, & T. J. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (pp. 9–49). San Francisco, CA: Jossey-Bass.
Milanowski, A. (2017). Lower Performance Evaluation Practice Ratings for Teachers of Disadvantaged Students. *AERA Open*, *3*(1), 1–16.
Rinehart, J. S., & Young, I. P. (1996). Effects of teacher gender and principal gender on ratings of teacher performance. *Journal of Personnel Evaluation in Education*, *10*(4), 313–323.
Rink, J. E. (2013). Measuring teacher effectiveness in physical education. *Research Quarterly for Exercise and Sport*, *84*(4), 407–418.
Rothstein, J. (2009). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, *4*(4), 537–571.
Salazar, M. d. C. (2018). Interrogating Teacher Evaluation: Unveiling Whiteness as the Normative Center and Moving the Margins. *Journal of Teacher Education*, *69*(5), 463–476.
Steinberg, M. P., & Garrett, R. (2016). Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measures? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317.
Sun, M. (2018). Black Teachers' Retention and Transfer Patterns in North Carolina: How Do Patterns Vary by Teacher Effectiveness, Subject, and School Conditions? *AERA Open*, *4*(3), 1–23.
U.S. Department of Education. (2016). *The State of Racial Diversity In the Educator Workforce* (Tech. Rep.). Washington, D.C.: U.S. Department of Education.
Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating Teachers with Classroom Observations Lessons Learned in Four Districts* (Tech. Rep.). Washington, D.C.: Brown Center on Education Policy at Brookings.

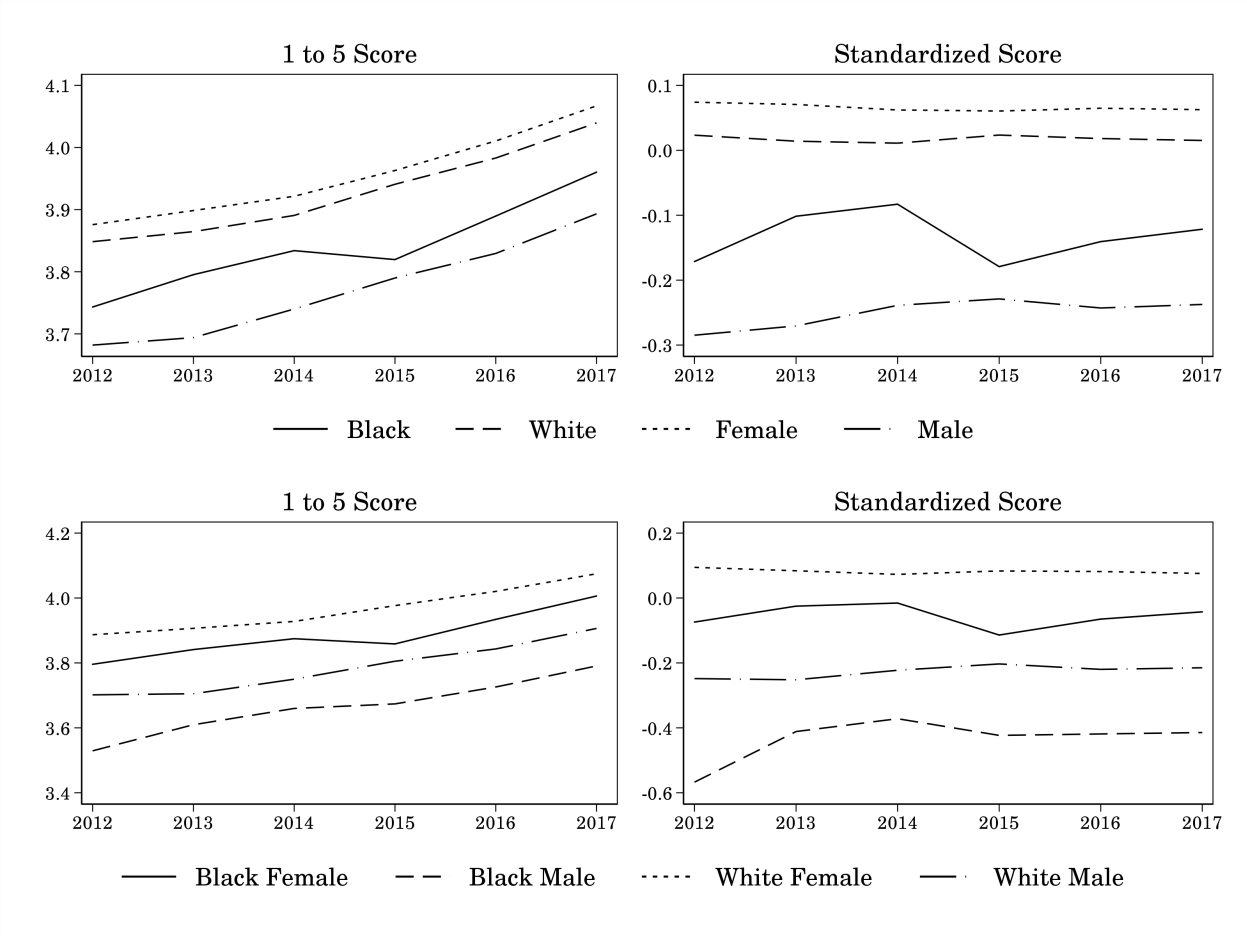

Figure 1. Average Observation Scores by Race and Gender

Notes: Each plot shows the average observation score across years for the subgroup defined in the plot legend. The plots of the left show the unadjusted scores, which range from 1 to 5. The plots on the right show scores that have been standardized within year.

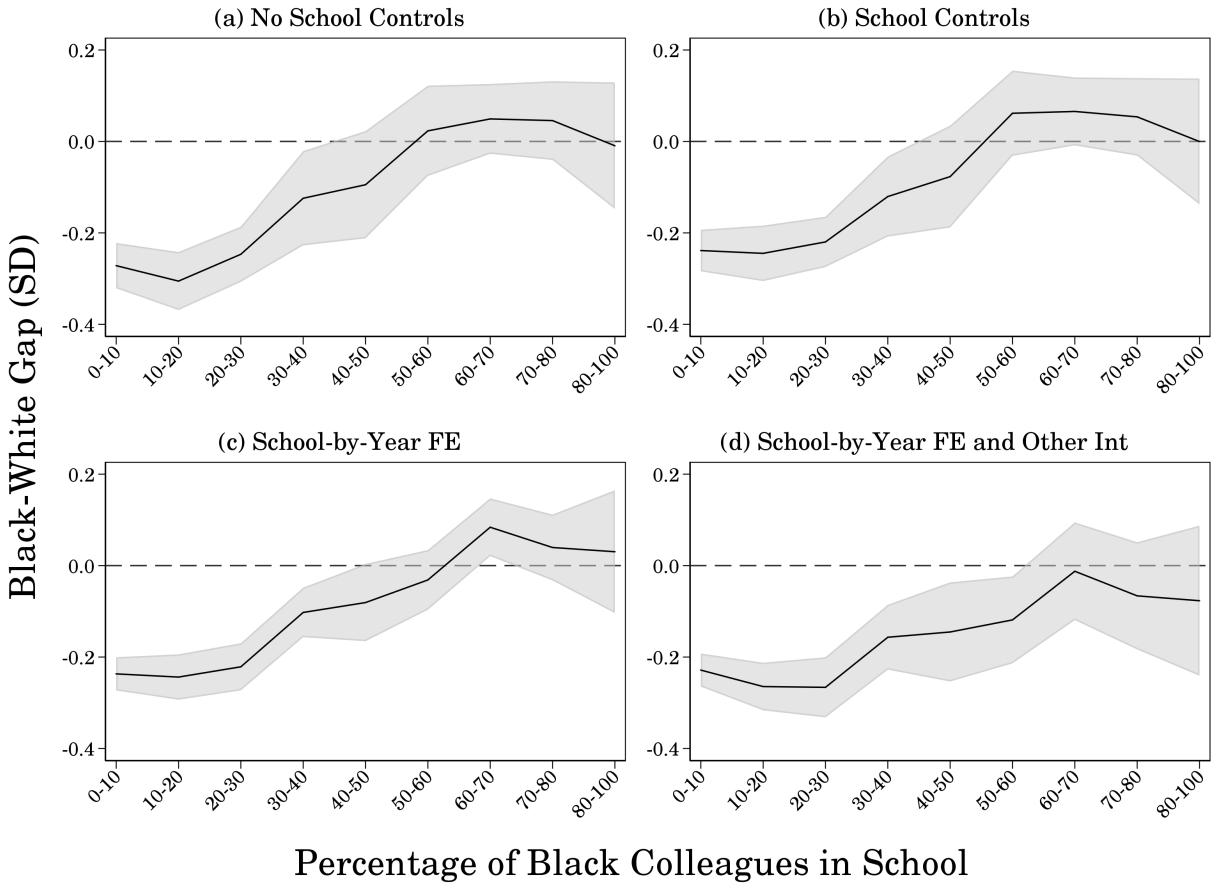

Figure 2. Black-White Gaps in Observation Scores by Teacher Racial Composition in School

Notes: Each plot shows the estimated contrast between Black and white teachers (i.e., the linear combination of the main effect of Black teacher and the interaction between Black teacher and the proportion of Black colleagues in the school) from a regression model that includes a categorical variable for the percentage of Black teachers in the school, not counting the focal teacher. All models include teacher demographic controls. Panels a and b include year fixed effects. Panel c includes school-by-year fixed effects, and panel d adds interactions between Black teacher and proportion of Black students and Black administrators, respectively. Shaded regions show 95% confidence intervals.

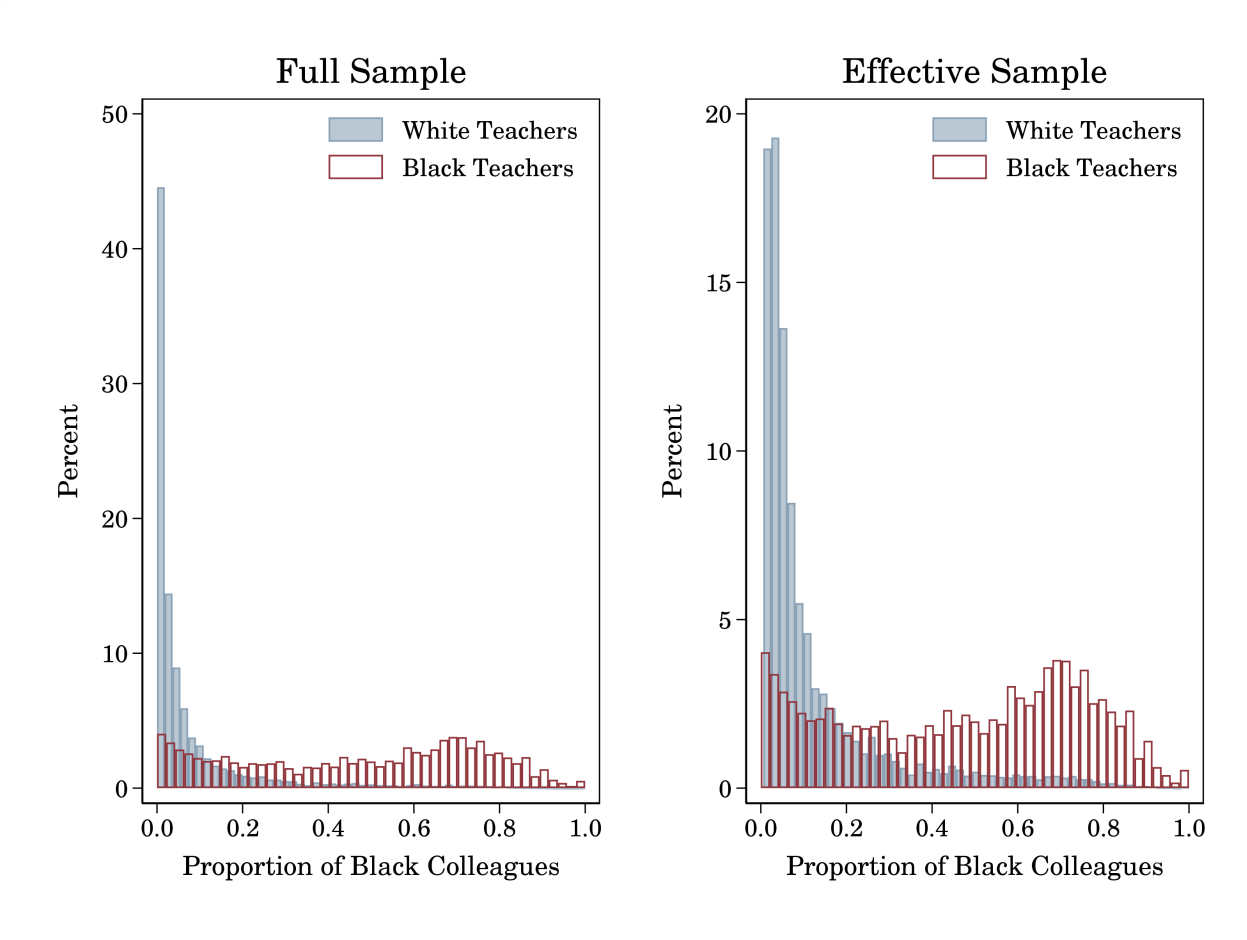

Figure 3. Distributions of Colleague Race

Notes: Each plot shows histograms for the proportion of Black colleagues in a teacher's school, separately for Black and white teachers. The y-axis indicates the percentage of teachers within a given racial group, such that the blue and red bars sum to 100 percent, respectively. The left plot shows the distributions for all teachers in the state. The right plot shows the distributions for teachers in the effective sample, which is defined by being in a school that has at least one Black teacher.

Table 1
Average Teacher, School, Colleague, and Observation Characteristics by Race and Gender

	All	Black	White	Female	Male
Teacher Characteristics					
Black	0.12			0.12	0.11
White	0.88			0.88	0.89
Female	0.79	0.80	0.79		
Male	0.21	0.20	0.21		
Age	42.6	43.9	42.5	42.6	42.8
Years of Experience	12.1	12.2	12.0	12.2	11.5
MA Degree	0.42	0.43	0.42	0.43	0.39
MA+ Degree	0.07	0.15	0.06	0.07	0.07
EdS Degree	0.07	0.10	0.07	0.07	0.07
PhD Degree	0.01	0.01	0.01	0.01	0.01
School Characteristics					
Enrollment (100s)	8.37	8.51	8.35	7.93	10.06
Prop. Black Students	0.23	0.64	0.17	0.23	0.23
Prop. Hispanic Students	0.08	0.11	0.08	0.08	0.07
Prop. Gifted Students	0.02	0.02	0.02	0.02	0.02
Prop. SPED Students	0.15	0.14	0.15	0.15	0.14
Prop. FRPL Students	0.57	0.76	0.55	0.58	0.55
Elementary School	0.52	0.50	0.52	0.60	0.20
Middle School	0.18	0.23	0.18	0.17	0.22
High School	0.26	0.24	0.26	0.19	0.51
Other School	0.04	0.03	0.04	0.03	0.06
Urban School	0.30	0.73	0.25	0.30	0.30
Suburban School	0.20	0.10	0.21	0.20	0.19
Town School	0.17	0.06	0.18	0.17	0.17
Rural School	0.33	0.11	0.36	0.33	0.34
Colleague Characteristics					
Prop. Black Colleagues	0.12	0.45	0.07	0.11	0.12
Prop. Male Colleagues	0.21	0.21	0.21	0.18	0.31
Black Principal	0.16	0.61	0.11	0.17	0.16
Male Principal	0.46	0.43	0.47	0.43	0.59
Prop. Black Administrators	0.18	0.62	0.13	0.18	0.18
Prop. Male Administrators	0.42	0.38	0.43	0.39	0.53
Observation Characteristics					
Average Observation Score (1 to 5)	3.92	3.84	3.93	3.96	3.77
Average Observation Score (SD)	0.00	-0.13	0.02	0.07	-0.25
Total Classroom Observations in Year	3.1	3.1	3.1	3.1	3.2
Total Raters in Year	1.9	1.9	1.9	1.8	2.0
N (Teacher-Year)	366783	42859	323924	290433	76350

Notes: Sample includes all Tennessee teachers with non-missing average observation scores from 2011–12 to 2016–17. Due to the very small number, we also drop non-Black, non-white teachers from the analysis.

Table 2
Gaps in Standardized Observation Scores by Subgroups

	(A) Race + Gender		(B) Race × Gender		
	Black	Male	Black Female	White Male	Black Male
School Level					
Elementary	-0.113***	-0.248***	-0.111***	-0.245***	-0.380***
Middle	-0.181***	-0.295***	-0.164***	-0.286***	-0.517***
High	-0.208***	-0.290***	-0.182***	-0.283***	-0.532***
School Locale					
Urban	-0.023	-0.261***	0.002	-0.226***	-0.353***
Suburban	-0.140***	-0.323***	-0.124**	-0.318***	-0.523***
Town/Rural	-0.351***	-0.345***	-0.336***	-0.342***	-0.739***
Observation Rubric					
TEAM	-0.386***	-0.306***	-0.373***	-0.302***	-0.732***
COACH	-0.226***	-0.352***	-0.202***	-0.343***	-0.642***
TEM	-0.179***	-0.332***	-0.184***	-0.347***	-0.505***
TIGER	-0.273***	-0.290***	-0.284***	-0.292***	-0.529***
Rubric Domain (TEAM)					
Instruction	-0.400***	-0.277***	-0.391***	-0.275***	-0.706***
Environment	-0.355***	-0.290***	-0.342***	-0.286***	-0.688***
Planning	-0.401***	-0.329***	-0.403***	-0.330***	-0.722***
Professionalism	-0.346***	-0.288***	-0.345***	-0.288***	-0.639***
Subject Taught					
Math	-0.153***	-0.418***	-0.125***	-0.406***	-0.645***
ELA	-0.205***	-0.297***	-0.186***	-0.281***	-0.645***
Science	-0.190***	-0.306***	-0.171***	-0.298***	-0.546***
Social Studies	-0.324***	-0.285***	-0.259***	-0.268***	-0.712***
Self-Contained	-0.135***	-0.400***	-0.135***	-0.400***	-0.530***
Foreign Language	-0.236***	-0.283***	-0.206**	-0.270***	-0.608***
Career/Tech Ed	-0.230***	-0.384***	-0.192***	-0.375***	-0.667***
Arts/Music	-0.194***	-0.108***	-0.137**	-0.093***	-0.365***
Health/P.E.	-0.045	-0.271***	-0.101*	-0.283***	-0.289***
Years of Experience					
0–1 Years	-0.176***	-0.185***	-0.177***	-0.186***	-0.358***
2–4 Years	-0.149***	-0.233***	-0.145***	-0.232***	-0.393***
5–20 Years	-0.131***	-0.307***	-0.115***	-0.298***	-0.495***
21+ Years	-0.281***	-0.417***	-0.279***	-0.415***	-0.708***

Notes: Each combination of row and panel (A and B) shows results from a separate regression model, where the row defines the subsample. For both panels, the reference group is white female. School-level clustered standard errors shown in parentheses. Observation scores are standardized within each year. For subject taught, subsamples include only teachers whose teaching assignment was 50% or more of the given subject. For rubric domain, we compute the yearly average within each teacher-by-year cell then standardize these teacher-by-year average scores. Rubric domain scores only include teachers evaluated using the TEAM rubric in 2012–13 to 2016–17. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3
Estimating Race and Gender Gaps in Classroom Observation Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Black Teacher	-0.150***	-0.183***	-0.043**	-0.158***	-0.360***	-0.286***
	(0.023)	(0.022)	(0.018)	(0.012)	(0.032)	(0.020)
Male Teacher	-0.317***	-0.293***	-0.257***	-0.260***	-0.254***	-0.258***
	(0.014)	(0.013)	(0.008)	(0.007)	(0.008)	(0.007)
Black Teacher x Prop. Black Students					0.587***	0.253***
					(0.061)	(0.035)
Teacher Characteristics						
MA Degree		0.124***	0.122***	0.121***	0.122***	0.120***
		(0.007)	(0.007)	(0.005)	(0.007)	(0.005)
MA+ Degree		0.217***	0.221***	0.134***	0.218***	0.133***
		(0.017)	(0.015)	(0.011)	(0.015)	(0.011)
EdS Degree		0.209***	0.229***	0.175***	0.222***	0.175***
		(0.017)	(0.016)	(0.010)	(0.016)	(0.010)
PhD Degree		0.305***	0.326***	0.281***	0.330***	0.282***
		(0.034)	(0.032)	(0.027)	(0.032)	(0.027)
Age 30–39		0.027***	0.023***	0.023***	0.023***	0.024***
		(0.009)	(0.009)	(0.007)	(0.009)	(0.007)
Age 40–49		-0.027**	-0.033***	-0.041***	-0.035***	-0.041***
		(0.011)	(0.010)	(0.009)	(0.010)	(0.009)
Age 50–59		-0.146***	-0.139***	-0.145***	-0.138***	-0.144***
		(0.014)	(0.013)	(0.011)	(0.013)	(0.011)
Age 60 and above		-0.288***	-0.276***	-0.281***	-0.274***	-0.280***
		(0.018)	(0.017)	(0.014)	(0.017)	(0.014)
Exp 0–4 years		-0.504***	-0.474***	-0.431***	-0.472***	-0.430***
		(0.012)	(0.011)	(0.009)	(0.011)	(0.009)
Exp 5–14 years		-0.090***	-0.081***	-0.077***	-0.082***	-0.077***
		(0.008)	(0.008)	(0.007)	(0.008)	(0.007)
Exp 25–39 years		0.133***	0.125***	0.118***	0.128***	0.118***
		(0.012)	(0.012)	(0.010)	(0.012)	(0.010)
Exp 40+ years		0.176***	0.173***	0.170***	0.173***	0.170***
		(0.039)	(0.038)	(0.033)	(0.038)	(0.033)
School Characteristics						
Enrollment (100s)			0.009**		0.009***	
			(0.003)		(0.003)	
Prop. Black Students			-0.150***		-0.277***	
			(0.056)		(0.059)	
Prop. Hispanic Students			-0.435***		-0.289**	
			(0.134)		(0.135)	
Prop. Gifted Students			1.834***		2.015***	
			(0.435)		(0.440)	
Prop. SPED Students			-0.114		-0.074	
			(0.205)		(0.203)	
Prop. FRPL Students			-0.315***		-0.336***	
			(0.055)		(0.055)	
Middle School			-0.180***		-0.179***	
			(0.030)		(0.030)	
High School			-0.185***		-0.185***	
			(0.035)		(0.035)	
Other School			-0.072		-0.075	
			(0.067)		(0.067)	
Urban School			0.017		0.020	
			(0.040)		(0.040)	
Town School			0.049		0.056	
			(0.039)		(0.039)	
Suburban School			-0.025		-0.021	
			(0.035)		(0.035)	
School-by-Year FE				✓		✓
N	355920	355920	355920	355920	355920	355920
R^2	0.019	0.073	0.096	0.361	0.099	0.362

Notes: In each model, the dependent variable is a teacher's average observation score in the given year. Scores are standardized within year. Models estimated via OLS. School-level clustered standard errors shown in parentheses. Models without school-by-year FE include year FE. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4
Disentangling Black Students, Black Teachers, and Black Administrators

	(1)	(2)	(3)	(4)	(5)	(6)
Black Teacher	-0.290***	-0.279***	-0.266***	-0.289***	-0.273***	-0.283***
	(0.020)	(0.017)	(0.018)	(0.020)	(0.020)	(0.020)
Interactions						
Black Tch. x Prop. Black Students	0.289***			-0.015	-0.006	0.017
	(0.035)			(0.053)	(0.052)	(0.053)
Black Tch. x Prop. Black Colleagues		0.451***		0.345***		
		(0.044)		(0.084)		
Black Tch. x Prop. Black Admin			0.262***	0.106**	0.111***	0.107**
			(0.029)	(0.043)	(0.042)	(0.044)
Black Tch. x Prop. Black Colleagues (Same Grade)					0.123***	0.103**
					(0.040)	(0.050)
Black Tch. x Prop. Black Colleagues (Other Grades)					0.220***	0.214***
					(0.077)	(0.080)
School-by-Year FE	✓	✓	✓	✓	✓	
School-by-Grade-by-Year FE						✓
N	351041	351041	351041	351041	351041	349787
R^2	0.352	0.353	0.352	0.353	0.353	0.472

Notes: In each model, the dependent variable is a teacher's average observation score in the given year. Scores are standardized within year. Models estimated via OLS. School-level clustered standard errors shown in parentheses. All models include the full vector of teacher controls. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 5
Do Observation Score Gaps Reflect Differences in Value-Added?

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black-White Gap	-0.170***	-0.213***	-0.131***	-0.166***	-0.176***	-0.164***	-0.151***	-0.151***
	(0.028)	(0.027)	(0.026)	(0.026)	(0.021)	(0.019)	(0.019)	(0.018)
Male-Female Gap	-0.310***	-0.255***	-0.260***	-0.243***	-0.281***	-0.248***	-0.246***	-0.236***
	(0.014)	(0.013)	(0.013)	(0.013)	(0.012)	(0.011)	(0.010)	(0.010)
Drift-Adjusted VA		0.306***		0.176***		0.298***		0.167***
		(0.006)		(0.007)		(0.006)		(0.005)
TVAAS Index			0.327***	0.237***			0.314***	0.235***
			(0.006)	(0.007)			(0.006)	(0.006)
Teacher Controls	✓	✓	✓	✓	✓	✓	✓	✓
School-by-Year FE					✓	✓	✓	✓
N	100989	100989	100989	100989	100989	100989	100989	100989
R^2	0.063	0.157	0.178	0.200	0.403	0.469	0.489	0.505

Notes: In each model, the dependent variable is a teacher's average observation score in the given year. Scores are standardized within year. Models estimated via OLS. Black-white and male-female Gaps are the estimated contrasts from the model. For male-female, this is equivalent to the regression coefficient for male teacher. For Black-white, it is equal to $\beta_1 BlackTch + \beta_2(BlackTch \times BlackColl)$, where the proportion of Black colleagues is set to 0.2, which is the mean of teachers in the effective sample. Drift-adjusted VA and TVAAS are standardized. School-level clustered standard errors shown in parentheses. All models include the full vector of teacher controls. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 6
Within-School Gaps in Student Assignment by Teacher Race and Gender

	Student Demographics					Prior-year Outcomes				
	Female	Black	FRPL	Gifted	SPED	ISS	OSS	Abs	Math	ELA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Black Teacher	-0.005**	0.029***	0.024***	-0.005***	0.018***	0.010***	0.008***	0.017***	-0.086***	-0.088***
	(0.002)	(0.002)	(0.002)	(0.001)	(0.003)	(0.001)	(0.001)	(0.004)	(0.010)	(0.010)
Black Tch. x Prop. Black Tch.	-0.003	-0.022***	-0.027***	0.004*	-0.016**	-0.002	0.008***	0.005	0.070***	0.060***
	(0.005)	(0.004)	(0.004)	(0.002)	(0.007)	(0.003)	(0.003)	(0.010)	(0.019)	(0.020)
Male Teacher	-0.047***	-0.001	-0.004***	0.001***	-0.004***	0.010***	0.006***	-0.002	0.016***	0.010***
	(0.002)	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)	(0.003)	(0.003)
School-by-Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Within-School SD	0.097	0.052	0.088	0.023	0.111	0.044	0.039	0.237	0.333	0.331
N	308681	308654	308681	308681	308681	307703	307703	304250	213015	212981
R^2	0.120	0.955	0.865	0.405	0.147	0.738	0.762	0.455	0.537	0.551

Notes: In each model, the dependent variable is the teacher-by-year mean of the student characteristic listed in the column header. In column 1, for instance, the dependent variable the proportion of a teacher's assigned students that are female. Student demographics are all expressed as proportions. For prior-year outcomes, ISS (in-school suspension) and OSS (out-of-school suspension) are the proportions of a teacher's assigned students who had at least one suspension of the given type in the prior school year. Absences, math achievement, and ELA achievement are the mean standardized prior-year scores for a teacher's assigned students. Models estimated via OLS. Sample restricted to teachers with subject/grade assignment data. School-level clustered standard errors shown in parentheses. All models include the full vector of teacher controls. The within-school standard deviation is calculated in two steps. First, we compute the standard deviation of each student assignment outcome within each school-by-year cell using all of the teachers in that school and year. Then, we compute the mean of these standard deviations across the full set of school-by-year cells. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7
Do Within-School Student Assignments Explain Teacher Observation Score Gaps?

			Achievement Sample		
	(1)	(2)	(3)	(4)	(5)
Black-White Gap	-0.185***	-0.149***	-0.181***	-0.142***	-0.141***
	(0.014)	(0.013)	(0.016)	(0.015)	(0.015)
Male-Female Gap	-0.248***	-0.239***	-0.264***	-0.254***	-0.257***
	(0.008)	(0.007)	(0.008)	(0.008)	(0.008)
Assigned Student Characteristics					
Prop. Female Students		0.203***		0.148***	0.125***
		(0.021)		(0.026)	(0.026)
Prop. Amer Ind Students		-0.517***		-0.174	-0.161
		(0.174)		(0.300)	(0.299)
Prop. Asian Students		0.160		0.509***	0.387***
		(0.104)		(0.155)	(0.149)
Prop. Black Students		-0.400***		-0.402***	-0.328***
		(0.048)		(0.057)	(0.057)
Prop. Hispanic Students		0.063		-0.016	-0.008
		(0.051)		(0.074)	(0.074)
Prop. Pac Isl Students		-0.307		-0.729*	-0.765*
		(0.234)		(0.425)	(0.421)
Prop. FRPL Students		-0.904***		-0.767***	-0.617***
		(0.026)		(0.039)	(0.039)
Prop. ELL Students		-0.151***		-0.052	0.142*
		(0.052)		(0.080)	(0.082)
Prop. Gifted Students		0.890***		0.550***	0.399***
		(0.107)		(0.092)	(0.088)
Prop. SPED Students		0.114***		0.004	0.173***
		(0.021)		(0.027)	(0.030)
Prop. Prior-year ISS		-0.284***		-0.245***	-0.194***
		(0.050)		(0.054)	(0.054)
Prop. Prior-year OSS		-0.152**		-0.358***	-0.305***
		(0.074)		(0.091)	(0.090)
Prop. Prior-year Expel		-1.084***		-1.039***	-1.028***
		(0.350)		(0.378)	(0.375)
Prop. Prior-year Retain		-0.220***		-0.780***	-0.507***
		(0.067)		(0.160)	(0.154)
Prior-year Absences (std)		-0.050***		-0.094***	-0.072***
		(0.008)		(0.017)	(0.016)
Prior-year Math (std)					0.101***
					(0.014)
Prior-year ELA (std)					0.055***
					(0.013)
Teacher Controls	✓	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓	✓
N	287335	287335	199030	199030	199030
R^2	0.380	0.394	0.391	0.406	0.407

Notes: In each model, the dependent variable is a teacher's average observation score in the given year. Scores are standardized within year. Models estimated via OLS. Black-white and male-female Gaps are the estimated contrasts from the model. For male-female, this is equivalent to the regression coefficient for male teacher. For Black-white, it is equal to $\beta_1 BlackTch + \beta_2(BlackTch \times BlackColl)$, where the proportion of Black colleagues is set to 0.2, which is the mean of teachers in the effective sample. School-level clustered standard errors shown in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 8
Do Subject and Grade Assignments Explain Observation Score Gaps?

	(1)	(2)	(3)	(4)
Black-White Gap	-0.149***	-0.146***	-0.149***	-0.147***
	(0.013)	(0.013)	(0.013)	(0.013)
Male-Female Gap	-0.239***	-0.232***	-0.250***	-0.237***
	(0.007)	(0.008)	(0.007)	(0.008)
Subject Taught (Proportion)				
Math		0.029***		0.029***
		(0.011)		(0.011)
ELA		ref.		ref.
Science		-0.102***		-0.105***
		(0.012)		(0.012)
Social Studies		-0.187***		-0.189***
		(0.013)		(0.013)
Self-Contained		-0.186***		-0.068***
		(0.012)		(0.013)
Foreign Language		-0.161***		-0.155***
		(0.020)		(0.020)
Career/Tech Ed		-0.012		0.001
		(0.015)		(0.015)
Arts/Music		0.011		0.044***
		(0.014)		(0.013)
Health/P.E.		-0.061***		-0.032**
		(0.015)		(0.015)
Grade Taught (Proportion)				
Pre-K			0.024	0.056
			(0.063)	(0.064)
Kindergarten			-0.109**	-0.081*
			(0.044)	(0.044)
Grade 1			-0.111**	-0.085*
			(0.043)	(0.044)
Grade 2			-0.219***	-0.193***
			(0.043)	(0.043)
Grade 3			-0.004	0.019
			(0.043)	(0.043)
Grade 4			0.078*	0.098**
			(0.043)	(0.043)
Grade 5			0.119***	0.134***
			(0.042)	(0.042)
Grade 6			0.002	0.015
			(0.039)	(0.039)
Grade 7			0.018	0.031
			(0.038)	(0.038)
Grade 8			0.113***	0.127***
			(0.038)	(0.038)
Grade 9			ref.	ref.
Grade 10			-0.049**	-0.044**
			(0.021)	(0.021)
Grade 11			0.015	0.051**
			(0.020)	(0.021)
Grade 12			-0.045**	-0.048**
			(0.022)	(0.023)
Teacher Controls	✓	✓	✓	✓
Assigned Student Controls	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓
N	287335	287335	287335	287335
R^2	0.394	0.399	0.400	0.403

Notes: In each model, the dependent variable is a teacher's average observation score in the given year. Scores are standardized within year. Models estimated via OLS. Black-white and male-female Gaps are the estimated contrasts from the model. For male-female, this is equivalent to the regression coefficient for male teacher. For Black-white, it is equal to $\beta_1 BlackTch + \beta_2(BlackTch \times BlackColl)$, where the proportion of Black colleagues is set to 0.2, which is the mean of teachers in the effective sample. Subject taught and grade taught, respectively, add up to a full assignment (proportion = 1.0) for each teacher. School-level clustered standard errors shown in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 9
Do Rater Characteristics Explain Observation Score Gaps?

	(1)	(2)	(3)	(4)	(5)
Black-White Gap	-0.145***	-0.145***	-0.125***	-0.121***	-0.107***
	(0.014)	(0.014)	(0.013)	(0.013)	(0.013)
Male-Female Gap	-0.206***	-0.206***	-0.186***	-0.185***	-0.185***
	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)
Rater Characteristics					
Black		0.010	0.010		
		(0.016)	(0.016)		
Male		0.024**	0.024**		
		(0.010)	(0.010)		
Ed.S. Degree		-0.000	-0.000	-0.007	-0.008
		(0.012)	(0.012)	(0.030)	(0.030)
Ph.D. Degree		-0.045***	-0.034**	0.035	0.035
		(0.016)	(0.016)	(0.045)	(0.045)
Assistant Principal		-0.057***	0.026***	0.023	0.024
		(0.011)	(0.010)	(0.021)	(0.021)
Teacher		-0.064***	0.066***	0.002	0.002
		(0.018)	(0.017)	(0.034)	(0.034)
Central Office		-0.326***	-0.158***	-0.109***	-0.109***
		(0.027)	(0.025)	(0.039)	(0.039)
3–5 Years Admin Exp.		0.015	0.010	-0.038***	-0.038***
		(0.010)	(0.009)	(0.011)	(0.011)
6–9 Years Admin Exp.		0.017	0.011	-0.064***	-0.064***
		(0.013)	(0.012)	(0.020)	(0.020)
10+ Years Admin Exp.		0.047***	0.047***	-0.060**	-0.060**
		(0.015)	(0.014)	(0.029)	(0.029)
Race Match w/ Teacher					0.033***
					(0.010)
Gender Match w/ Teacher					-0.003
					(0.005)
Observation Order					
Second			0.151***	0.152***	0.152***
			(0.005)	(0.004)	(0.004)
Third			0.366***	0.366***	0.366***
			(0.009)	(0.008)	(0.008)
Fourth			0.436***	0.435***	0.435***
			(0.011)	(0.010)	(0.010)
Fifth or more			0.585***	0.582***	0.582***
			(0.016)	(0.015)	(0.015)
Total Observations					
Three			-0.506***	-0.502***	-0.502***
			(0.008)	(0.008)	(0.008)
Four			-0.867***	-0.855***	-0.855***
			(0.014)	(0.013)	(0.013)
Five or more			-1.008***	-0.999***	-0.999***
			(0.013)	(0.012)	(0.012)
School-by-Year FE	✓	✓	✓	✓	✓
Rater FE				✓	✓
N	559977	559977	559977	559773	559773
R^2	0.295	0.298	0.367	0.407	0.407

Notes: In each model, the dependent variable is a teacher's average item-level score for a given observation, where teachers have multiple observations in each year. Scores are standardized within year. Models estimated via OLS. Black-white and male-female Gaps are the estimated contrasts from the model. For male-female, this is equivalent to the regression coefficient for male teacher. For Black-white, it is equal to $\beta_1 BlackTch + \beta_2(BlackTch \times BlackColl)$, where the proportion of Black colleagues is set to 0.15, which is the mean of teachers in the effective sample. All models include the full vector of teacher, assigned student, and subject/grade controls. School-level clustered standard errors shown in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 10
Observation Score Gaps by Teacher-Principal Work History

		Var = Black		Var = Male	
	(1)	(2)	(3)	(4)	(5)
Black-White Gap	-0.130***	-0.156***	-0.127***	-0.130***	-0.130***
	(0.015)	(0.021)	(0.021)	(0.015)	(0.015)
Male-Female Gap	-0.195***	-0.195***	-0.195***	-0.200***	-0.192***
	(0.009)	(0.009)	(0.009)	(0.011)	(0.012)
Did Principal Hire Teacher?					
P Hired T		0.004		0.006	
		(0.010)		(0.011)	
Var x P Hired T		0.052**		0.011	
		(0.025)		(0.014)	
Time Working Together					
1–2 Years			0.059***		0.058***
			(0.019)		(0.019)
3–5 Years			0.075***		0.075***
			(0.022)		(0.022)
6–9 Years			0.100***		0.106***
			(0.024)		(0.024)
10+ Years			0.101***		0.110***
			(0.030)		(0.030)
Var x 1–2 Years			-0.011		0.003
			(0.023)		(0.013)
Var x 3–5 Years			-0.007		-0.003
			(0.030)		(0.017)
Var x 6–9 Years			0.008		-0.032
			(0.042)		(0.022)
Var x 10+ Years			0.092		-0.026
			(0.063)		(0.041)
School-by-Year FE	✓	✓	✓	✓	✓
N	283657	283657	283657	283657	283657
R^2	0.417	0.417	0.417	0.417	0.417

Notes: Interactions are defined by the "Var" listed in the column header. "P Hired T" is a binary indicator equal to 1 if the principal entered the school in the same year or earlier than the teacher. Sample only includes observations performed by school principals. In each model, the dependent variable is a teacher's average item-level score for a given observation, where teachers have multiple observations in each year. Scores are standardized within year. Models estimated via OLS. Black-white and male-female Gaps are the estimated contrasts from the model. For male-female, this is equivalent to the regression coefficient for male teacher. For Black-white, it is equal to $\beta_1 BlackTch + \beta_2(BlackTch \times BlackColl)$, where the proportion of Black colleagues is set to 0.15, which is the mean of teachers in the effective sample. All models include the following controls: teacher characteristics, assigned student characteristics, subject/grade assignment, observer characteristics, observation order, and total number of observations. School-level clustered standard errors shown in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 11
Observation Score Gaps by Observation Order

		Var = Black		Var = Male	
	(1)	(2)	(3)	(4)	(5)
Black-White Gap	-0.122***	-0.080***	-0.080***	-0.122***	-0.121***
	(0.013)	(0.015)	(0.014)	(0.013)	(0.013)
Male-Female Gap	-0.185***	-0.185***	-0.185***	-0.165***	-0.168***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Observation Order					
Second	0.152***	0.156***	0.120***	0.154***	0.115***
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Third	0.367***	0.373***	0.299***	0.381***	0.301***
	(0.008)	(0.009)	(0.010)	(0.009)	(0.010)
Fourth	0.435***	0.438***	0.346***	0.440***	0.340***
	(0.010)	(0.010)	(0.012)	(0.010)	(0.012)
Fifth or more	0.582***	0.588***	0.457***	0.585***	0.445***
	(0.015)	(0.015)	(0.018)	(0.015)	(0.018)
Var x Second		-0.053***	-0.065***	-0.012**	-0.004
		(0.011)	(0.012)	(0.006)	(0.007)
Var x Third		-0.080***	-0.102***	-0.060***	-0.044***
		(0.016)	(0.018)	(0.009)	(0.011)
Var x Fourth		-0.033	-0.074***	-0.022*	0.001
		(0.023)	(0.026)	(0.013)	(0.016)
Var x Fifth or more		-0.072**	-0.108***	-0.011	0.013
		(0.031)	(0.039)	(0.016)	(0.021)
Within-Rater Observation Order					
Second			0.086***		0.093***
			(0.006)		(0.006)
Third			0.131***		0.136***
			(0.011)		(0.011)
Fourth			0.150***		0.167***
			(0.016)		(0.017)
Fifth or more			0.236***		0.239***
			(0.025)		(0.026)
Var x Second			0.028**		-0.019**
			(0.014)		(0.009)
Var x Third			0.055**		-0.007
			(0.026)		(0.017)
Var x Fourth			0.100**		-0.035
			(0.043)		(0.025)
Var x Fifth or more			-0.007		-0.021
			(0.062)		(0.038)
School-by-Year FE	✓	✓	✓	✓	✓
Rater FE	✓	✓	✓	✓	✓
N	559271	559271	559271	559271	559271
R^2	0.407	0.407	0.408	0.407	0.408

Notes: Interactions are defined by the "Var" listed in the column header. In each model, the dependent variable is a teacher's average item-level score for a given observation, where teachers have multiple observations in each year. Scores are standardized within year. Models estimated via OLS. Black-white and male-female Gaps are the estimated contrasts from the model. For male-female, this is equivalent to the regression coefficient for male teacher. For Black-white, it is equal to $\beta_1 BlackTch + \beta_2(BlackTch \times BlackColl)$, where the proportion of Black colleagues is set to 0.15, which is the mean of teachers in the effective sample. All models include the following controls: teacher characteristics, assigned student characteristics, subject/grade assignment, observer characteristics, and total number of observations. School-level clustered standard errors shown in parentheses * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 12
Teacher Attitudes Towards Evaluation System

	(1)	(2)	(3)	(4)
(A) The processes used to conduct my teacher evaluation are fair to me.				
Black Teacher	0.017	0.120***	0.155***	0.156***
	(0.018)	(0.019)	(0.017)	(0.018)
Black Teacher x Black-White Observation Gap			0.269***	0.054
			(0.036)	(0.037)
Male Teacher	0.040**	0.038**	0.083***	0.081***
	(0.017)	(0.016)	(0.018)	(0.018)
Male Teacher x Male-Female Observation Gap			0.179***	-0.035
			(0.041)	(0.040)
Observation Score (std)				0.218***
				(0.007)
(B) The teacher evaluation process has helped me improve my teaching.				
Black Teacher	0.220***	0.245***	0.265***	0.266***
	(0.017)	(0.019)	(0.018)	(0.018)
Black Teacher x Black-White Observation Gap			0.153***	0.033
			(0.034)	(0.035)
Male Teacher	-0.030*	0.033**	0.055***	0.053***
	(0.017)	(0.016)	(0.019)	(0.019)
Male Teacher x Male-Female Observation Gap			0.086**	-0.033
			(0.037)	(0.037)
Observation Score (std)				0.121***
				(0.007)
(C) The evaluation system is a burden.				
Black Teacher	-0.054*	-0.176***	-0.194***	-0.199***
	(0.028)	(0.037)	(0.035)	(0.034)
Black Teacher x Black-White Observation Gap			-0.141**	-0.007
			(0.057)	(0.058)
Male Teacher	-0.132***	-0.082***	-0.098***	-0.097***
	(0.024)	(0.028)	(0.028)	(0.028)
Male Teacher x Male-Female Observation Gap			-0.059	0.083
			(0.070)	(0.071)
Observation Score (std)				-0.140***
				(0.012)
Teacher Controls	✓	✓	✓	✓
School-by-Year FE		✓	✓	✓

Notes: Models estimated via OLS. In columns 1 and 2, school-level clustered standard errors are shown in parentheses. Columns 3 and 4 interact *Black Teacher* and *Male Teacher* with the estimated Black-white Gap and male-female Gap in observation scores, respectively, for the teacher's school-by-year cell. An increase in the Black-white or male-female Observation Gap indicates an increase in the observation scores of Black or male teachers relative to their white or female teachers in the same school and year. Columns 3 and 4 report school-level cluster bootstrapped standard errors in parentheses (500 repetitions). For each panel, the dependent variable is the standardized score for the listed survey item. The unit of observation is teacher-by-year. See footnote 17 for exact description of survey items. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 13
How Much of Observation Score Gaps Can We Explain?

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male-Female Gap	-0.243***	-0.228***	-0.223***	-0.226***	-0.215***	-0.230***	-0.211***	-0.184***
	(0.013)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)
Black-White Gap (0–25% Black Colleagues)	-0.222***	-0.149***	-0.159***	-0.126***	-0.141***	-0.116***	-0.140***	-0.100***
	(0.031)	(0.024)	(0.023)	(0.024)	(0.024)	(0.024)	(0.023)	(0.022)
Black-White Gap (25-50% Black Colleagues)	-0.111***	-0.073**	-0.092***	-0.045	-0.078**	-0.063*	-0.079**	-0.062**
	(0.037)	(0.033)	(0.032)	(0.032)	(0.032)	(0.033)	(0.034)	(0.031)
Black-White Gap (50-100% Black Colleagues)	0.083**	0.072**	0.049	0.090**	0.084**	0.042	0.059*	0.038
	(0.039)	(0.036)	(0.035)	(0.036)	(0.036)	(0.039)	(0.034)	(0.035)
School-by-Year FE		✓	✓	✓	✓	✓	✓	✓
Teacher Characteristics			✓					✓
Assigned Student Characteristics				✓				✓
Subject/Grade Assignment					✓			✓
Rater Characteristics						✓		✓
Value-Added							✓	✓
N	280274	280274	280274	280274	280274	280025	280274	280025
R^2	0.167	0.367	0.373	0.376	0.378	0.414	0.389	0.455

Notes: In each model, the dependent variable is a teacher's average item-level score for a given observation, where teachers have multiple observations in each year. Scores are standardized within year. Models estimated via OLS. The male-female gap is the estimated coefficient for male teacher. We estimate the Black-white gap for teachers in schools with 0–25%, 25–50%, and 50–100% Black colleagues, respectively, by interacting an indicator for Black teacher with an indicator for the given group. All models include controls for the main effect of colleague race, observation order, and the total number of observations received in that year. Rater characteristics includes time-varying characteristics, rater fixed effects, and binary indicators for race and gender match. For value-added we use the drift-adjusted estimates rather than TVAAS to maximize sample size. Columns 6 and 8 differ in sample size due to dropping of singleton observations. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Appendix
Online Appendix

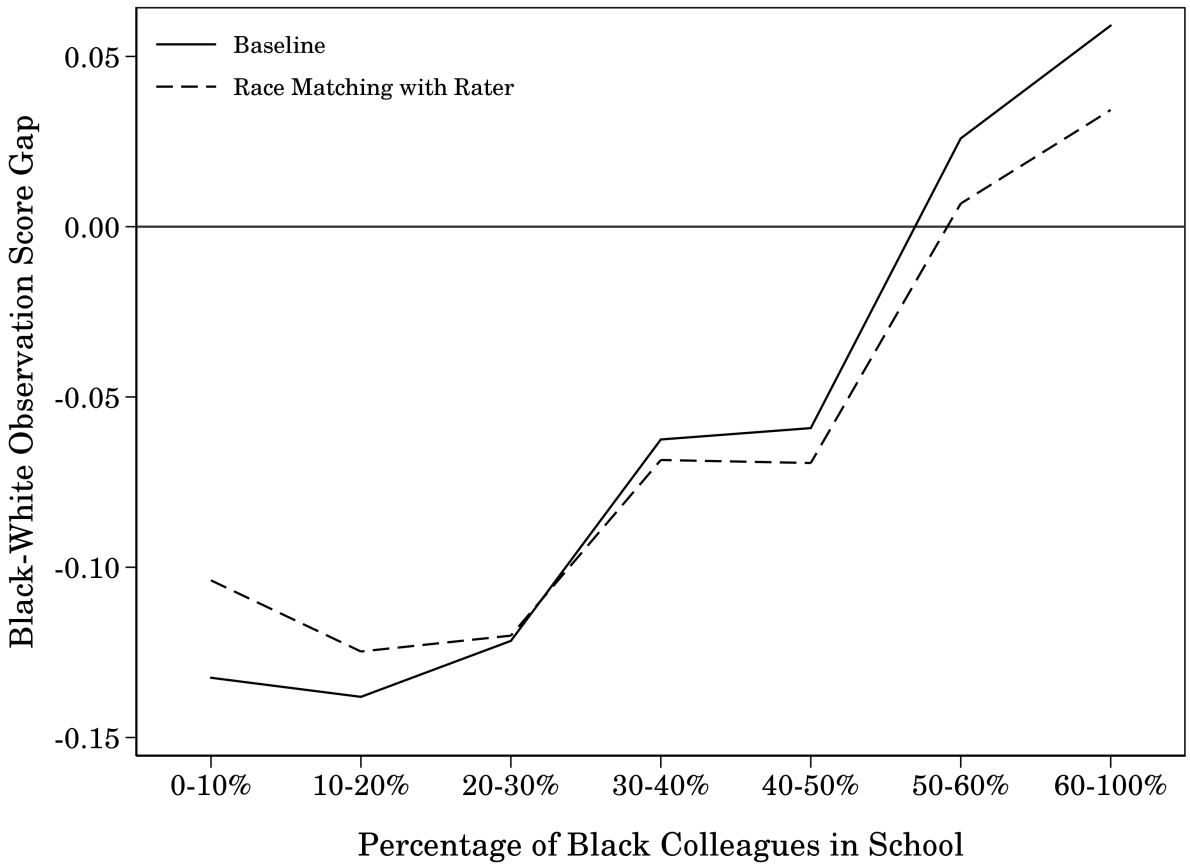

Figure A1. Black-White Observation Score Gap with and without Controlling for Teacher-Rater Race Match

Notes: This plot shows the estimated Black-white gap (i.e., the linear combination of the main effect for Black teacher and the interaction between Black teacher and proportion of Black colleagues in the school) in observation scores from a model that includes school-by-year FE, rater FE, teacher characteristics, assigned student characteristics, subject/grade assignment, observation order, total number of observations, rater characteristics, and the interaction between Black teacher and a categorical variable for the proportion of Black colleagues in the school. This model corresponds to the results shown in Table 9 column 5, except that the proportion of Black colleagues variable is categorical instead of continuous.

Table A1

The Relationship Between Observation Scores and Teacher and School Characteristics by Race and Gender Subgroups

	Subgroup			
	Black (1)	White (2)	Female (3)	Male (4)
Teacher Characteristics				
MA Degree	0.126***	0.130***	0.127***	0.102***
	(0.020)	(0.007)	(0.007)	(0.013)
MA+ Degree	0.173***	0.242***	0.233***	0.147***
	(0.028)	(0.017)	(0.016)	(0.030)
EdS Degree	0.247***	0.219***	0.223***	0.240***
	(0.033)	(0.018)	(0.017)	(0.030)
PhD Degree	0.298***	0.336***	0.342***	0.282***
	(0.069)	(0.035)	(0.037)	(0.056)
Age 30–39	0.027	0.006	0.015	0.022
	(0.030)	(0.009)	(0.010)	(0.018)
Age 40–49	-0.054	-0.049***	-0.039***	-0.044**
	(0.034)	(0.011)	(0.011)	(0.021)
Age 50–59	-0.244***	-0.136***	-0.149***	-0.135***
	(0.039)	(0.013)	(0.014)	(0.026)
Age 60 and above	-0.430***	-0.268***	-0.281***	-0.285***
	(0.048)	(0.017)	(0.019)	(0.032)
Exp 0–4 years	-0.500***	-0.477***	-0.506***	-0.367***
	(0.030)	(0.011)	(0.012)	(0.020)
Exp 5–14 years	-0.068***	-0.086***	-0.099***	-0.011
	(0.024)	(0.008)	(0.009)	(0.017)
Exp 25–39 years	0.112***	0.134***	0.135***	0.079***
	(0.033)	(0.012)	(0.013)	(0.028)
Exp 40+ years	0.306***	0.147***	0.181***	0.156**
	(0.108)	(0.042)	(0.044)	(0.074)
School Characteristics				
Enrollment (100s)	0.012**	0.009**	0.006*	0.015***
	(0.005)	(0.004)	(0.003)	(0.004)
Prop. Black Students	0.404***	-0.253***	-0.163***	-0.209***
	(0.097)	(0.059)	(0.054)	(0.070)
Prop. Hispanic Students	0.220	-0.370***	-0.344**	-0.861***
	(0.236)	(0.129)	(0.134)	(0.185)
Prop. Gifted Students	3.702***	1.663***	2.036***	1.000*
	(0.531)	(0.447)	(0.434)	(0.590)
Prop. SPED Students	0.576	-0.138	-0.104	-0.171
	(0.401)	(0.194)	(0.186)	(0.345)
Prop. FRPL Students	-0.347***	-0.362***	-0.325***	-0.310***
	(0.114)	(0.055)	(0.055)	(0.079)
Middle School	-0.282***	-0.215***	-0.171***	-0.244***
	(0.056)	(0.031)	(0.030)	(0.037)
High School	-0.265***	-0.275***	-0.153***	-0.299***
	(0.060)	(0.036)	(0.036)	(0.040)
Other School	-0.074	-0.148**	-0.029	-0.214***
	(0.111)	(0.068)	(0.064)	(0.078)
Urban School	0.029	0.012	-0.001	0.066
	(0.067)	(0.041)	(0.040)	(0.055)
Town School	0.015	0.065*	0.061	0.008
	(0.085)	(0.039)	(0.038)	(0.057)
Suburban School	-0.103	-0.016	-0.026	-0.027
	(0.076)	(0.035)	(0.034)	(0.051)
N	41353	314567	281921	73999
R^2	0.088	0.090	0.086	0.073

Notes: In each model, the dependent variable is a teacher's average observation score in the given year and the sample is defined by the subgroup listed in the column header. Scores are standardized within year. Models estimated via OLS. School-level clustered standard errors shown in parentheses. Models without school-by-year FE include year FE. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A2

Within-School Gaps in Student Assignment by Race and Gender, Conditioning on Proportion of Black Students

	Student Demographics				Prior-year Outcomes				
	Female	FRPL	Gifted	SPED	ISS	OSS	Abs	Math	ELA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Black Teacher	-0.005***	0.009***	-0.002***	0.007***	0.005***	0.006***	0.009***	-0.033***	-0.038***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.006)	(0.006)
Male Teacher	-0.047***	-0.004***	0.001***	-0.004***	0.011***	0.007***	-0.002	0.014***	0.008***
	(0.002)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)	(0.003)	(0.003)
Prop. Black Stu (Tch)	-0.024**	0.357***	-0.085***	0.289***	0.171***	0.165***	0.341***	-1.892***	-1.838***
	(0.010)	(0.022)	(0.012)	(0.020)	(0.014)	(0.012)	(0.029)	(0.107)	(0.108)
Black Tch. x Prop. Black Tch	-0.003	-0.019***	0.002	-0.010	0.002	0.012***	0.012	0.027	0.019
	(0.005)	(0.003)	(0.002)	(0.007)	(0.003)	(0.003)	(0.010)	(0.018)	(0.019)
School-by-Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
N	308654	308654	308654	308654	307674	307674	304223	213008	212974
R^2	0.121	0.872	0.415	0.163	0.746	0.773	0.459	0.570	0.582

Notes: In each model, the dependent variable is the teacher-by-year mean of the student characteristic listed in the column header. In column 1, for instance, the dependent variable the proportion of a teacher's assigned students that are female. Student demographics are all expressed as proportions. For prior-year outcomes, ISS (in-school suspension) and OSS (out-of-school suspension) are the proportions of a teacher's assigned students who had at least one suspension of the given type in the prior school year. Absences, math achievement, and ELA achievement are the mean standardized prior-year scores for a teacher's assigned students. Models estimated via OLS. Sample restricted to teachers with subject/grade assignment data. School-level clustered standard errors shown in parentheses. All models include the full vector of teacher controls. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A3

Within-School Gaps in Student Assignment by Race and Gender, Conditioning on Proportion of Female Students

	Student Demographics				Prior-year Outcomes				
	Black	FRPL	Gifted	SPED	ISS	OSS	Abs	Math	ELA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Black Teacher	0.025***	0.018***	-0.004***	0.013***	0.009***	0.009***	0.017***	-0.066***	-0.069***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.007)	(0.007)
Male Teacher	-0.001**	-0.009***	0.002***	-0.020***	0.005***	0.002***	-0.010***	0.058***	0.060***
	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)	(0.003)	(0.003)
Prop. Female Stu (Tch)	-0.008**	-0.103***	0.018***	-0.333***	-0.111***	-0.096***	-0.177***	0.806***	0.966***
	(0.004)	(0.006)	(0.002)	(0.008)	(0.005)	(0.004)	(0.013)	(0.024)	(0.023)
Black Tch. x Prop. Black Tch	-0.022***	-0.027***	0.004*	-0.017***	-0.002	0.008***	0.005	0.074***	0.065***
	(0.004)	(0.004)	(0.002)	(0.006)	(0.003)	(0.003)	(0.010)	(0.018)	(0.019)
School-by-Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
N	308654	308681	308681	308681	307701	307701	304250	213014	212980
R^2	0.955	0.866	0.406	0.207	0.748	0.772	0.458	0.562	0.587

Notes: In each model, the dependent variable is the teacher-by-year mean of the student characteristic listed in the column header. In column 1, for instance, the dependent variable the proportion of a teacher's assigned students that are Black. Student demographics are all expressed as proportions. For prior-year outcomes, ISS (in-school suspension) and OSS (out-of-school suspension) are the proportions of a teacher's assigned students who had at least one suspension of the given type in the prior school year. Absences, math achievement, and ELA achievement are the mean standardized prior-year scores for a teacher's assigned students. Models estimated via OLS. Sample restricted to teachers with subject/grade assignment data. School-level clustered standard errors shown in parentheses. All models include the full vector of teacher controls. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4
Race and Gender Matching with Rater

	% of Black Colleagues in School					
	0–10%		10–30%		30–100%	
	(1)	(2)	(3)	(4)	(5)	(6)
Black Teacher	-0.127***	-0.120***	-0.127***	-0.123***	-0.029	-0.027
	(0.029)	(0.030)	(0.020)	(0.020)	(0.024)	(0.025)
Male Teacher	-0.174***	-0.173***	-0.216***	-0.215***	-0.250***	-0.253***
	(0.008)	(0.008)	(0.015)	(0.015)	(0.024)	(0.023)
Black Rater	0.077**		0.025		-0.029	
	(0.031)		(0.025)		(0.027)	
Male Rater	0.015		0.032		0.075***	
	(0.012)		(0.022)		(0.027)	
Race Match w/ Teacher	0.025	0.031	0.044***	0.043***	0.038**	0.040**
	(0.022)	(0.024)	(0.016)	(0.016)	(0.018)	(0.019)
Gender Match w/ Teacher	-0.005	-0.003	-0.005	-0.004	-0.001	-0.013
	(0.006)	(0.005)	(0.014)	(0.012)	(0.017)	(0.017)
School-by-Year FE	✓	✓	✓	✓	✓	✓
Rater FE		✓		✓		✓
N	436634	436475	79573	79461	43708	43593
R^2	0.362	0.402	0.355	0.396	0.358	0.407

Notes: In each model, the dependent variable is a teacher's average item-level score for a given observation, where teachers have multiple observations in each year. Scores are standardized within year. Models estimated via OLS. The sample is defined by the percentage of Black colleagues in the school according to the range listed in the column header. All models include the full vector of teacher, assigned student, rater, and subject/grade controls. School-level clustered standard errors shown in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A5

How Much of Observation Score Gaps Can We Explain? (Including Teachers Without Individual Value-Added

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Male-Female Gap	-0.222***	-0.210***	-0.204***	-0.202***	-0.200***	-0.212***	-0.184***
	(0.012)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)
Black-White Gap (0–25% Black Colleagues)	-0.233***	-0.160***	-0.169***	-0.134***	-0.156***	-0.134***	-0.100***
	(0.023)	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)	(0.022)
Black-White Gap (25-50% Black Colleagues)	-0.127***	-0.086***	-0.104***	-0.059**	-0.088***	-0.084***	-0.062**
	(0.033)	(0.028)	(0.027)	(0.027)	(0.027)	(0.027)	(0.031)
Black-White Gap (50-100% Black Colleagues)	0.049	0.048*	0.025	0.059**	0.051**	0.029	0.038
	(0.031)	(0.025)	(0.025)	(0.025)	(0.025)	(0.027)	(0.035)
School-by-Year FE		✓	✓	✓	✓	✓	✓
Teacher Characteristics			✓				✓
Assigned Student Characteristics				✓			✓
Subject/Grade Assignment					✓		✓
Rater Characteristics						✓	✓
N	559945	559945	559945	559945	559945	559742	280025
R^2	0.160	0.347	0.353	0.356	0.353	0.388	0.455

Notes: In each model, the dependent variable is a teacher's average item-level score for a given observation, where teachers have multiple observations in each year. Scores are standardized within year. Models estimated via OLS. The male-female gap is the estimated coefficient for male teacher. We estimate the Black-white gap for teachers in schools with 0–25%, 25–50%, and 50–100% Black colleagues, respectively, by interacting an indicator for Black teacher with an indicator for the given group. All models include controls for the main effect of colleague race, observation order, and the total number of observations received in that year. Rater characteristics includes time-varying characteristics, rater fixed effects, and binary indicators for race and gender match. Columns 6 and 7 differ in sample size due to dropping of singleton observations. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.