Potential Race and Gender Biases in High-Stakes Teacher Observations

Jason A. Grissom

Vanderbilt University

Brendan Bartanen

University of Virginia

Abstract

Rubric-based classroom observation ratings make up the largest component of summative evaluation ratings given to teachers in the multiple-measure evaluation systems states have implemented in the last decade. Using data from the first eight years of statewide implementation of teacher evaluation in Tennessee, we document race and gender gaps in observation ratings and evaluate the extent to which these gaps reflect true differences in instructional effectiveness. Descriptively, white and female teachers outscore their Black and male colleagues by 0.15 SD and 0.30 SD, on average. Gaps persist even when accounting for other measures of teachers' effectiveness, such as value-added to student test scores or student attendance, which we interpret as evidence consistent with bias. We then investigate sources of these potential biases. We document that the Black–white gap is largest in schools where Black teachers are racially isolated and is partially explained by Black teachers' propensity to be assigned less advantaged students within their schools. We also find evidence that teachers receive somewhat higher ratings from raters of the same race. In contrast, we find no same-gender rater effects, and in fact, beyond some differences by subjects and grades taught, uncover few explanations for the large advantage women see in observation ratings.

Potential 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 systematically reflect factors other than their job performance. As Cohen and Goldhaber (2016) point out, there are multiple potential sources of bias, including rater subjectivity and contextual factors beyond the teacher's control. As an example of this latter source, studies have found that teachers assigned lower-performing students are given lower observation ratings (e.g., Whitehurst, Chingos, & Lindquist, 2014), including in a sample from the Measures of Effective Teaching (MET) project in which students were randomly assigned to teachers (Steinberg & Garrett, 2016), which suggests that factors other than lower instructional quality in classrooms with lower-achieving students can lead to lower ratings. While a relatively large literature has examined accuracy in teacher value-added or related student growth metrics that often comprise some portion of teachers' summative ratings (e.g.,

Chetty, Friedman, & Rockoff, 2014; Kane & Staiger, 2008; Rothstein, 2009), little research has examined such issues in ratings from the classroom observation that make up the largest component of the overall evaluation in most systems (Grissom & Youngs, 2016).

This study investigates whether classroom observation ratings demonstrate bias with respect to teacher race and gender. That is, we look for evidence that a teacher's race or gender predicts systematic departures between the observation ratings they are assigned and the teaching performance that observations purport to measure. To be clear, our definition of bias includes mechanisms that may arise from multiple sources. Raters may unconsciously (or consciously) apply observation rubrics differently to teachers with different characteristics (Grissom & Loeb, 2017). The rubrics themselves may devalue teaching practices that teachers from some subgroups are more likely to use (e.g., teachers of color employing culturally relevant pedagogy) relative to other practices that are similarly effective (Salazar, 2018). Rubrics similarly may leave out or assign low scores to teaching behaviors that are more effective with some student subgroups, such as low-achieving students or students with disabilities (Jones, 2016; Milanowski, 2017). Studies suggest that women and teachers of color are more likely to teach such students (Kalogrides, Loeb, & Béteille, 2013), illustrating one mechanism whereby nonrandom sorting of students and teachers across classrooms might contribute to race and gender bias in observation scores.

We look for evidence of racial and gender bias in classroom observation scores in Tennessee over the first eight years of the implementation of the state's high-stakes, multiple-measure teacher evaluation system (2011–12 to 2018–19). Tennessee's system mandates that classroom observations be scored with an approved, standards-based observation rubric by a trained rater following a codified set of observation procedures. Observation ratings are combined with test score-based measures to determine teachers' overall evaluation ratings. We first ask whether classroom observation ratings differ by teacher race or gender. To examine potential bias, we then analyze these ratings in a

regression framework that allows us to test for gaps by race or gender even after accounting for alternative measures of teacher performance, such as teachers' value-added to student test scores or student attendance. Documenting that racial and gender gaps in scores persist, we turn to factors that may contribute to these gaps, including school context, characteristics of teaching assignments within schools, and characteristics of raters, each of which may differ systematically for teachers from different demographic backgrounds.

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 (and their outcomes), subject taught, and other characteristics of the school environment. Our analysis is based on data from approximately 460,000 teacher-by-year observations.

We uncover large gaps in classroom observation ratings by both teacher gender and race. These gaps change little over the eight-year study period. Pooling across years, women outscore men by 0.30 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, with some variation 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 ratings as a function of teacher value-added to test scores and student attendance, plus other potential proxies for teacher effectiveness (e.g., experience level, degree attainment), we still find relatively large rating advantages for female and white teachers. We interpret this evidence as consistent with gender and race bias in ratings, as we define it. To be explicit, our interpretation of bias rests on the assumption that these alternative measures and proxies account for any true differences in instructional effectiveness according to teacher race and gender. However, we implement a bounding exercise to demonstrate that any unobserved dimensions of

effectiveness would have to be much more strongly related to teacher race or gender than observable dimensions to explain these gaps.

Delving into the various mechanisms behind these potential biases, we find that 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. We also find that Black teachers' scores are lower when they are more racially isolated; that is, their scores are lower when they have few Black colleagues, while the gap with white teachers disappears or even reverses in schools that have a majority of Black teachers. Moreover, leveraging variation within school and 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, we have less success explaining the likely bias in favor of women in observation ratings, beyond observing that men are more likely to teach grades and subjects where scores are lower, on average. We find no evidence that teachers benefit from being observed by a rater of the same gender.

This study extends a small body of existing research on racial/ethnic and gender gaps in classroom observation ratings. In a study of MET data, Campbell and Ronfeldt (2018) find 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 ratings in MET were low-stakes, and raters' behavior in low- and high-stakes settings can differ substantially (Grissom & Loeb, 2017), so investigating these patterns in the case in which evaluation scores can be used for personnel decisions remains important. Three more recent studies have taken this step. First, Drake, Auletto, and Cowen (2019) examine summative ratings assigned to teachers in Michigan during the initial years of that

state's implementation of its new evaluation system. They show that male teachers and teachers of color were given lower ratings, and that low ratings of teachers of color were more common in schools with higher proportions of white teachers. As the authors note, however, ratings in Michigan 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. Second, (Steinberg & Sartain, 2020) investigate racial gaps in elementary teachers' classroom observation ratings for two years of Chicago's REACH system. Like other studies, they find that Black teachers receive lower scores than white teachers, a pattern that persists after accounting for test score value-added. However, they find that Black–white rating differences are driven almost completely by between-school differences in ratings; the gap becomes indistinguishable from zero after comparing teachers within schools. Moreover, they uncover no evidence of a "race match" effect for teachers and evaluators. Our study builds on (Steinberg & Sartain, 2020) by examining both racial and gender differences in ratings among teachers across grade spans in a statewide system with wide variation in locale type and school and faculty composition. Perhaps due to differences in setting and modeling approaches, we reach different conclusions about the nature of observation score gaps than these prior studies. Finally, Chi (2021) examines high-stakes teacher observations from a large school district in North Carolina, finding evidence of positive teacher–rater match effects for race and gender. While our race-match estimate is nearly identical in magnitude, we do not find evidence of a gender match effect.

Conceptualizing Bias in Teacher Observation Ratings

We define bias as the systematic deviation of teacher observation scores from actual instructional effectiveness. To formalize this idea, we define θ_{it} to be true instructional effectiveness for teacher i in year t. Instructional effectiveness is the combination of a teacher's skill and effort towards improving student outcomes relevant to schools, such as academic learning and social-emotional skills. This value is unobserved. Instead, we observe a measure of instructional effectiveness, $\hat{\theta}_{it}$, in the form of a classroom observation rating. Observation rubrics aim to measure instructional effectiveness by defining a set of teaching practices and behaviors that are purported to lead to improved student outcomes. Raters then apply the rubric when observing a teacher's lesson to produce a classroom observation score.

We can decompose $\hat{\theta}$ as follows:

$$\hat{\theta}_{it} = \theta_{it} + \delta_{it} + \epsilon_{it} \tag{1}$$

where ϵ_{it} is measurement error that is orthogonal to θ_{it} and δ_{it}. For instance, ϵ_{it} captures non-persistent factors that affect observation scores, such as being observed on an "off day" or having an observation during the teacher's best-behaved class section. The parameter δ_{it}, then, captures any persistent factors that lead to differences between true and measured effectiveness. This definition is similar to the definition of "teacher-level bias" in value-added models proposed by Rothstein (2009) and subsequently discussed by Chetty et al. (2014). In the context of classroom observation scores, δ_{it} is a function of potentially many factors, only some of which have been documented in prior research. An example is school or classroom context; teachers who teach more students with low baseline achievement, students from low-income families, and students of color receive lower observation scores (Steinberg & Garrett, 2016). Other potential sources of bias include the rubric employed to assess classroom practice, which may define instructional effectiveness

inappropriately or too narrowly, or the rater assigned to the teacher, who may not apply the rubric correctly, given the complex cognitive demands of mapping instruction onto multiple indicators (Cohen & Goldhaber, 2016; Milanowski, 2017).

The question we ask is whether δ_{it} is a function of teacher race or gender. For ease of exposition, we simplify to the case of two groups within each racial or gender classification, and employ parallel definitions of racial and gender bias as follows:

$$\delta_{Group} = E[\delta_{it}|Group_i = A] - E[\delta_{it}|Group_i = B] \tag{2}$$

That is, we define racial (gender) bias as the difference in average systematic departures from true instructional effectiveness for Black and white (male and female) teachers. Given the conceptualization of δ_{it}, we emphasize that our definition of race and gender bias encompasses mechanisms beyond simple group member discrimination. Notably, it includes systemic processes (i.e., any component of δ_{it}) that may differentially affect teachers' ratings by their race and gender. For example, we interpret as bias a situation in which male teachers are systematically assigned a higher proportion of students with behavior challenges, and facing more behavior challenges lowers a teacher's observation ratings.

We do not observe δ_{Group}. Instead, we observe $\hat{\theta}_{Group}$, which are mean differences in observation scores by race and gender. Based on equation 1, these observation score differences can be decomposed as follows:

$$\hat{\theta}_{Group} = \theta_{Group} + \delta_{Group} \tag{3}$$

In other words, differences in average ratings we observe for Black and white teachers (or male and female teacher) conflate racial (or gender) bias and any average differences in true instructional effectiveness between groups. Such differences might arise if, for example, differences in propensity to turn over mean that one group has higher average experience

than the other, and more experience makes a teacher more effective. The potential for such differences means that we cannot interpret the descriptive gap, $\hat{\theta}_{Group}$, as evidence of bias.

Our first empirical challenge thus becomes attempting to isolate δ_{Group}, that is, to assess whether race or gender bias in observation ratings exist and to measure its magnitude. Assuming these values can be isolated, our second empirical task becomes identifying the components of δ_i that are the drivers of δ_{Group}.

Data

This study analyzes administrative data from Tennessee, a state made up of 147 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.[1] 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 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. In all systems, teachers' overall summative evaluation scores (called the "level of effectiveness") comprise the weighted average of three components: scores from formal classroom observations, student test score growth (measured by TVAAS, the state's value-added metric), and an alternative measure of student achievement.[2] Our analysis

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

[2] For teachers in tested classrooms, observations are weighted 50% and TVAAS 35%. Approximately 40% of teachers are assigned individual TVAAS scores based on the performance of the students in their

focuses on classroom observation scores, which receive the greatest weight in determining teachers' summative ratings.

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 whether the educator is identified as male or female (binary) and their race/ethnicity classification (white, Black, Hispanic, Asian, Native American, or other). In Tennessee, the fraction of Asian, Native American, and other race/ethnicity educators is 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 2018–19. 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 with scores on individual observation indicators and rater identifiers for teachers in districts using the TEAM observation rubric, which are 82% of the state's teachers. Additionally, in 2015–16 and 2016–17, we can access observation-level information for teachers from Shelby County Schools (one of the districts that uses an alternative observation rubric), which is important because it is the district with the largest number of Black teachers in the state.

Classroom observation ratings are assigned by raters using the rubric associated with their evaluation system (e.g., TEAM, COACH). Raters are required by the state to

---

classrooms on tests measuring subjects they teach. Historically, the remainder of teachers have been assigned the school's overall TVAAS score for this component, though in more recent years the state has identified alternative growth metrics based on portfolios, for example, and reduced the weight given to growth in favor of more weight to observations. Much of our analysis is limited to teachers with individual TVAAS scores. The achievement metric, which is locally chosen, is omitted from our analysis.

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

complete a training and certification to conduct observations. 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] This rubric was based on the National Institute for Excellence in Teaching (NIET) Teaching Standards Rubric, adapted in a collaboration between NIET and TDOE. It is generic in the sense that it is used for observing all teachers, regardless of subject or specialization.[5] 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,[6] 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.

Similar to national trends, Tennessee's teacher workforce is far less racially diverse than the student population and overwhelmingly female. Eighty-nine percent of teachers are white and 79% are female.[7] While Black and white teachers in Tennessee have similar

[4] According to TDOE guidelines, the first three domains are scored based on formal classroom observations, while the professionalism scores may be assigned outside the context of a classroom observation. This distinction, alongside the fact that the indicators in the professionalism domain (e.g., leadership) relate more to work outside a teacher's classroom, suggests that this component may not be a good measure of a teacher's instructional effectiveness, though in practice professionalism scores are highly correlated with scores on the other domains. We look separately at domain-level scores below.

[5] The TEAM rubrics are available at https://team-tn.org/evaluation/teacher-evaluation-2/. Although the same rubric is used for all teachers, TDOE provides guidance on applying the rubric differently for some specialized teachers, such as special educators or those in career and technical education.

[6] State policy 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.

[7] Appendix Table A1 shows average teacher, school, colleague, and observation characteristics for Tennessee teachers, along with disaggregations by race and gender.

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 65% 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 demographics, the average white teacher works in a school with few Black teachers (7%) and is unlikely to have a Black principal (10%). While 64% of Black teachers work in a school with a Black principal, only 46% of their colleagues are Black, on average. Mainly due to the sorting by school level, men have more male colleagues and are more likely to work for a male principal.

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 3.96 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

Our analysis follows from the conceptual framework presented in section 2, where a teacher's observation score ($\hat{\theta}_{it}$) is a function of their true instructional effectiveness (θ_{it}) and both persistent (δ_{it}) and transient factors (ϵ_{it}) that lead to differences between true and measured effectiveness: $\hat{\theta}_{it} = \theta_{it} + \delta_{it} + \epsilon_{it}$. Race and gender gaps in observation scores ($\hat{\theta}_{Group}$) can then be decomposed into systematic biases (δ_{Group}) and true instructional effectiveness differences (θ_{Group}) between Black and white or male and female teachers. Repeated here for clarity, our definition of bias encompasses any elements of δ_{it} that are

correlated with teacher race or gender.

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

$$\hat{\theta}_{it} = \beta_0 + \beta_1 Black_i + \beta_2 Male_i + \epsilon_{it} \tag{4}$$

where the average observation score (standardized by year) of teacher i 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.[8] β_1 and β_2 are descriptive estimates of $\hat{\theta}_{Group}$ and should not be interpreted as clean estimates of race and gender bias because they include both θ_{it} and δ_{it}.

To better isolate δ_{it}, the systematic bias portion of these observation score gaps, we control for multiple determinants of or proxies for teachers' instructional effectiveness:

$$Score_{ist} = \alpha + \beta_1 Black_i + \beta_2 Male_i + \gamma X_{it} + \epsilon_{ist} \tag{5}$$

where X_{it} is a vector of two teacher characteristics, years of experience (entered categorically) and educational attainment, that may partially capture differences in teachers' actual effectiveness. More important, for teachers in tested grades and subjects, we can also include measures of "value added" (VA) to two outcomes for subsets of teachers. For teachers of tested grades and subjects, we estimate value added to student test scores to capture teachers' individual contributions to student achievement. For teachers in self-contained classrooms in grades 1–5, we estimate value added to student attendance. To construct each of these VA measures, we follow the leave-year-out, drift-adjusted approach outlined in Chetty et al. (2014).[9] Prior work demonstrates only a

[8] 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.

[9] The estimation steps are as follows. First, we residualize student test scores (separately by subject) on a

weak correlation between teachers' contributions to raising student test scores and their contributions to lowering student absenteeism (Gershenson, 2016; Liu & Loeb, 2019), suggesting that teacher effectiveness is a multi-dimensional construct. Including multiple alternative measures of effectiveness allows us to more convincingly isolate the portion of race and gender gaps that may be attributable to bias.

Equation 5 is equivalent to a version of the Kitagawa–Blinder–Oaxaca decomposition where γ is the non-discriminatory coefficient vector estimated via a pooled regression of both groups (Jann, 2008). β_1 and β_2 are differences in observation scores between Black and white and male and female teachers, respectively, conditional on the elements of X_{it}.[10] We can interpret these as estimates of race and gender bias, δ_{Group}, under the assumption that $E[\theta_{it}|X_{it}] = E[\theta_{it}|X_{it}, Group_i]$. That is, we must assume that X_{it} fully accounts for any instructional effectiveness differences along race and gender lines.

There are two types of threats to this assumption. First, there could exist dimensions of true instructional effectiveness that are orthogonal to X_{it} but are correlated with teacher race or gender. A teacher's ability to build such social-emotional skills as self-management or peer relationship development, for instance, might be poorly captured by their contributions to student test scores. If this dimension is then correlated with teacher race

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. Finally, we standardize this measure at the teacher level within each year. For attendance VA, we follow the same procedure except we do not control for prior-year test scores, since they are unavailable for students in early grades. Additionally, we restrict the sample to teachers in self-contained classrooms in grades 1–5. Because our data only measures daily attendance (as opposed to by each class period), we cannot plausibly isolate the contribution of teachers to student attendance when students have multiple teachers per day.

[10] Specifically, these parameters can be expressed as:
$\mathbb{E}(X_A)'(\gamma_A - \gamma) + \mathbb{E}(X_B)'(\gamma_B - \gamma) + (\alpha_A - \alpha) + (\alpha_B - \alpha)$, which shows that the unexplained differences in observation scores between groups A and B (i.e., Black and white or male and female teachers) reflect differences in the coefficient vectors for determinants of instructional effectiveness ($\gamma_A - \gamma$ and $\gamma_B - \gamma$) and differences in the group-specific intercepts ($\alpha_A - \alpha$ and $\alpha_B - \alpha$). In fact, we find few differences in the coefficient vectors, with the vast majority of unexplained differences driven by differences in the intercepts.

or gender, our estimate of δ_{Group} would be incorrect. We aim to address this threat by drawing on two distinct outcome-based measures of teacher effectiveness (test score and attendance VA), as well as measures derived from student surveys for a subsample of teachers. The second type of threat is that X_{it} includes all relevant measures of instructional effectiveness but contains measurement error, which leads to an understatement of θ_{Group} due to attenuation bias and a corresponding overstatement of δ_{Group}. This concern is salient for value-added measures, where reliability can be low (Koedel, Mihaly, & Rockoff, 2015). To mitigate the potential for this type of attenuation bias, we rely on value-added models that incorporate multiple years of test score data and use shrinkage to reduce measurement error. Our efforts to address these threats likely substantially reduce any residual race or gender gap attributable to differences in true job performance. Still, we also conduct a bounding exercise (described in Appendix C) to examine the robustness of our results to violations of this key assumption.

Given estimates of race and gender bias, the second part of our analysis investigates potential mechanisms. That is, what are the elements of δ_{it} that are correlated with teacher race or gender? We first investigate the role of school context by adding to equation 5 a set of school characteristics, including average student demographics, enrollment size, school level, and school locale. Attenuation in the estimated race and gender gaps when controlling for school characteristics would suggest that school contextual factors play a role in producing these biases, though we cannot necessarily determine whether the observables are the relevant factors or whether they are proxies for unobservable factors. Given the limited set of school characteristics in administrative data, we can go a step further by replacing school characteristics with school-by-year fixed effects, which eliminates all between-school heterogeneity, thus isolating race and gender bias to within-school mechanisms. Beyond controlling for biases from between-school heterogeneity, these school-by-year fixed effects further assuage concerns that teacher characteristics insufficiently adjust for any true differences in instructional effectiveness.

We proceed to investigate a number of within-school mechanisms for racial and gender biases using a school-by-year fixed effects approach. These mechanisms include systematic differences by race and gender in students assigned, in teaching assignment (e.g., subject taught), and in rater characteristics. For students assigned, we add information about students in teacher's classroom to the model, including their background characteristics and prior-year test scores, attendance, and disciplinary information. For assignment, we examine a teacher's grade level and subject.[11] For rater characteristics, we leverage the fact that teachers have multiple observations over the course of the year, typically performed by multiple raters (most often the principal and an assistant principal). Rater characteristics include race, gender, education level, experience, and job title, though we also estimate specifications that include rater fixed effects. School-by-year and rater fixed effects are identified given that 91% of schools have multiple raters in a given year.

Descriptive Gaps in Observation Scores

We begin our analysis by descriptively examining 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 2012 is equal to 2019. The third pattern is that the male–female gap is

[11] Because teachers can have multiple subject assignments, we operationalize subject taught in proportional terms, with full (100%) ELA teachers as the reference category.

larger than the Black–white gap. Pooling across all years, women outscore men by 0.30 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.[12]

Table 1 shows descriptive gaps in teacher-by-year observation scores across six categories of subgroups: school level, school locale, the teacher observation rubric used by the district, the rubric domain, the teacher's primary subject taught[13], 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 teachers. 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) as in elementary schools (-0.12 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 (-0.37 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

[12] 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.

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

gaps for health/P.E., math, and self-contained teachers. The Black–white gap is largest among teachers with more than 20 years of experience (-0.25 SD) and brand-new teachers (-0.22 SD for 0–1 years of experience).

Gender gaps generally are less variable in magnitude across subgroups. For instance, the male–female gap is -0.23 SD, -0.29 SD, and -0.28 SD in elementary, middle, and high schools, respectively. Similar to the Black–white gap, the male–female gap is largest in town/rural schools (-0.33 SD), though there is also a sizable gap in urban schools (-0.25 SD). Subject taught and teacher experience show the greatest variability in the magnitude of the gender gap. The gap is largest for math and self-contained teachers (-0.39 SD) and smallest for arts/music teachers (-0.11 SD). The gender gap also steadily grows across the experience distribution, from -0.18 SD among first- and second-year teachers to -0.39 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 0.72 SD lower than white women, on average.

Do Observation Score Gaps Reflect Bias?

The previous section establishes that there are large differences in average observation scores along race and gender lines. Next, we attempt to isolate the portion of these gaps that are attributable to racial and gender biases by controlling for observable determinants of teacher job performance.

Table 2 shows estimated Black–white and male–female gaps in observation scores with and without controlling for teacher characteristics and/or teacher value-added measures. Each set of columns is a particular sample as we can only construct value-added estimates for a subset of teachers. The patterns are similar across each set of models; while

the male–female gap shrinks when accounting for teacher characteristics and value-added, the Black–white gap increases. For instance, Column 1 shows the baseline race (-0.15 SD) and gender gaps (-0.30 SD) for the full sample of teachers, while column 2 adds controls for teacher educational attainment and experience. In Tennessee, Black and female teachers are slightly more experienced and have slightly higher educational attainment than white and male teachers, respectively. Given that both experience and educational attainment are associated with higher observation scores, adjusting for these characteristics increases the estimated race gap and (slightly) decreases the gender gap.

Teacher experience and degree attainment may be quite limited as proxies for effectiveness. In the remaining columns of Table 2, we add more direct measures of effectiveness based on student outcomes as covariates. Before turning to these results, note that in terms of mean test score value-added, Black and white teachers are very similar, while men are somewhat lower-performing than women (-0.18 SD). For attendance value-added, Black teachers are higher-performing than white teachers (0.30 SD), on average, with no substantial differences by gender. As shown in columns 4 and 6, each of these value-added measures is positively associated with observation scores, though the relationship is much stronger for test score value-added. Column 8 also shows that they are independently predictive of observation scores, suggesting that test score and attendance value-added reflect different dimensions of performance that are both captured in observation scores.[14] Controlling for either value-added measure (in addition to teacher characteristics) further widens the Black–white gap and further narrows the male–female gap, though even in the latter case, the residual gap remains substantial. Column 8 shows residual gaps of -0.14 SD for Black teachers and -0.23 SD for male teachers, accounting for both VA measures simultaneously.

We interpret the patterns in Table 2 as evidence of race and gender bias in

[14] Similar to prior work (e.g., Gershenson, 2016; Liu & Loeb, 2019), we find only weak correlations between teachers' test score and attendance value-added ($r = 0.09$).

observation scores by our definition that bias is a systematic deviation of ratings from actual instructional effectiveness. This evidence is not definitive; interpreting the residual gaps in Table 2 as bias necessarily rests on assumptions that are not directly testable. Perhaps most obviously, we must assume that Black and white (male and female) teachers do not differ on an important dimension of instructional effectiveness that is not captured by observable characteristics or contributions to student test scores and attendance. This argument would be bolstered by consideration of additional alternative measures of effectiveness. For a small sample of schools, we can incorporate such an additional measure: ratings from student surveys. Student surveys, which allow districts to capture student perceptions of instructional quality and the classroom environment, are optional under the state's evaluation system, and teacher-level measures are available for only 5% of our sample.[15] As Appendix Table A2 shows, accounting for student survey scores does not eliminate the race or gender gap in this sample.

An alternative threat to the bias claim might be that summative observation ratings in Tennessee incorporate dimensions of teacher performance beyond instruction. In particular, teachers receive ratings in TEAM on a professionalism domain that measure contributions to the school outside the classroom (e.g., community involvement, teacher leadership). Because we have defined bias as deviation from instructional effectiveness, we may be concerned that gaps in professionalism scores drive overall gaps, or that teachers' performance on this component may not correlate well with degree, experience, or alternative performance measures generated from student outcomes. In contrast, in the absence of bias, we might expect that gaps on the domains most closely linked to teachers' instructional work, like *instruction* or *environment* (which captures expectations, classroom

[15] Under the state's educator evaluation system, districts may choose to administer one of a handful of approved student engagement surveys (Tennessee School Climate Survey, Tripod Survey, My Student Survey, or Panorama) to count for five percent of a teacher's summative evaluation rating. Unfortunately, we can only access the 1 to 5 rating that was assigned to a teacher according to these surveys—we cannot observe the survey employed or the criteria determining the score. Nonetheless, Appendix Table A2 shows a clear positive relationship between a teacher's average classroom observation score and their 1 to 5 student survey rating, even when controlling for their test score VA.

management, and classroom culture), would be drastically reduced when the covariates are included. To investigate, in Appendix Tables A3 and A4 we estimate models separately by rubric domain. The table shows that estimated gaps are large not just for professionalism but for the three domains more closely linked to teachers instructional work as well. Accounting for teacher characteristics and the VA measures, residual gaps remain large (and generally similar in magnitude) across all four domains, spanning those most clearly connected to the other effectiveness measures (e.g., *instruction*) to those that are less so.[16]

We consider two other robustness analyses to complement Table 2. First, Appendix Table A5 replicates the patterns in Table 2 using school-by-year fixed effects. While this eliminates potential biases arising between-school differences (an issue we return to in the next section), it may provide a more robust lower-bound estimate of race and gender bias by only comparing teachers working in the same school context. Indeed, results show smaller, though still significant, estimates of the racial gap, though estimates of the gender gap are very close to those shown in Table 2. Second, we re-estimate the test score value-added results using an alternative value-added metric—namely, the measure calculated by the state for the evaluation system (TVAAS). Appendix Table A6 shows that this measure has predictive value for observation ratings even while conditioning on the value-added score we estimated. The estimates of the racial and gender gaps, however, remain qualitatively unchanged when we account for this measure.[17]

[16] In fact, there is little evidence that item- or domain-level scores measure different constructs, particularly for the instruction, planning, and environment domains. When analyzed in an exploratory factor analysis, all items load strongly on a single latent construct. Professionalism shows some evidence of being a distinct construct, but if we predict factor scores from professionalism items only and, separately, all instruction, planning, and environment items, the two sets of scores are correlated at 0.64.

[17] 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, by construction, does not. 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

Finally, we perform a bounding exercise proposed by Oster (2019). The intuition of this check is to use the change in the estimated gaps with and without controls for observable determinants of teacher instructional effectiveness, along with the change in R^2, to compute adjusted gaps in the presence of unobservable differences in instructional effectiveness. The results of this exercise show that the residual race gap is robust even to a large unobserved difference in a facet of instructional effectiveness between Black and white teachers, while the residual gender gap is robust to a small or medium unobserved difference between male and female teachers (see Appendix C).

What Drives Race and Gender Biases in Observation Scores?

The remainder of our analysis focuses on examining potential drivers of the apparent systematic deviations in observation scores from teacher effectiveness along race and gender lines. We start by investigating the role of school context, then move to within-school mechanisms, such as rater characteristics and teacher-student assignment patterns. For these analyses, we focus on the sample of teachers for whom we can estimate test score VA, rather than both test score and attendance VA. Not controlling for attendance VA allows us to include roughly three times as many teachers, and once we condition on test score VA, further controlling for attendance VA leads to little or no change in the estimated gaps. For completeness, Appendix B shows results for two alternative samples: the full sample of teachers (including those without test score VA) and the sample of teachers with both attendance and test score VA.

Differences in School Context

Table 3 shows Black–white and male–female gap estimates with controls for school context. The estimates in column 1 show the baseline estimates, which only include teacher characteristics and test score VA. Column 2 adds school characteristics, including

predict teachers' current year performance, drift-adjusted VA will be predictive of observation scores even conditional on TVAAS.

enrollment size, student demographics, school level, and locale type. The average Black teacher and white teacher in Tennessee work in very different school contexts. For example, the average Black teacher works in a school where 65% of students are Black, compared to only 17% for the average white teacher (see Appendix Table A1). 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 that are unrelated to an individual teacher's own effectiveness. 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). Accounting for school characteristics, then, helps uncover the extent to which score biases associate with race are explained by teacher sorting patterns.

As shown in column 2, adding these controls reduces the Black–white gap from -0.21 to -0.03 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 decreases slightly in magnitude when accounting for school characteristics (-0.26 SD to -0.24 SD). The decrease is explained by a single factor: men are much more likely to work in high schools (52% to 19%), where teachers receive systematically lower observation scores.

Taken at face value, the estimates in column 2 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 across environments that vary in their average

observation scores. However, when we replace school characteristics with school-by-year fixed effects in column 3, we recover a Black–white gap of -0.13 SD.[18] The estimated male–female gap is very similar in columns 2 and 3.

What explains the large difference in the Black teacher coefficients for these specifications? Our reanalysis by race and gender subgroups (shown in Appendix Table A7) uncovers that the model in column 2 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 0.44 for Black teachers and -0.45 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* × *Proportion Black Students*) is substantial.[19]

Columns 4 and 5 of 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 4) or school-by-year fixed effects (column 5), the interaction term is positive, large in magnitude, and statistically significant. In substantive terms, the *Black Teacher* × *Prop. Black Students* coefficient in column 5 demonstrates that the Black–white gap in observation scores is largest in schools that have few Black students and smallest in

[18] In models with school-by-year FE, the effective sample for the Black–white gap conditions on having at least one Black teacher in the school. Descriptively, schools that have no Black teachers are overwhelmingly located in areas designated as town/rural (73% vs. 35% for schools with at least one Black teacher) and that have an average of 89% white students (vs. 52%). One potential reason for the change in the *Black Teacher* coefficient between columns 2 and 3, then, is the change in the effective sample. To check this, we re-estimated the model in column 2 but restricted to school-by-year cells that had at least one Black teacher. As shown in Appendix Table A8, the estimated Black–white gap was almost identical. Thus, the difference between columns 2 and 3 is not driven by a change in the effective sample.

[19] 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.

schools with many Black students.

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 0.35 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 that the positive interaction between Black teacher and proportion of Black students is 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.[20] 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 characteristics 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

[20] In Table 4, Column 4, the p-value for a test of the difference between the interaction terms for Black Teacher × Prop. Black Students and Black Teacher × Prop. Black Colleagues (Black Teacher × Prop. Black Admin) is $p = 0.11$ ($p = 0.10$).

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 -0.20 SD in schools with 0–30% Black colleagues to zero or even positive in schools with a majority of Black colleagues.[21]

The results in Tables 3 and 4 show that school context is an important moderator of the Black–white observation score gap and it is thus important to estimate a model that accounts for this heterogeneity. Instead of estimating a single parameter for Black teacher, then, we include an interaction to estimate the Black–white gap for teachers in schools with 0–25%, 25–50%, and 50–100% Black colleagues.[22] The results from this approach, shown in column 5, are equivalent to Figure 2 Panel C, except that we reduce the number of categories for parsimony.

---

[21] Appendix Figure A1 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.

[22] These categories, respectively, include 31%, 21%, and 48% (92%, 5%, 3%) of Black (white) teachers in the state. An alternative is to report both the main effect (*Black Teacher*) and interaction term (*Black Teacher* × *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 effects. 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.

Within-School Mechanisms

The previous section establishes that substantial race and gender gaps persist even when restricting comparisons to Black and white (male and female) teachers working in the same school in the same year. We now turn to investigating explanations for these residual within-school gaps.

First, we consider the characteristics of raters. Here, we leverage the fact that teachers receive multiple classroom observations each year, which often are conducted by different raters (typical principals and assistant principals). 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 5, column 1 shows the baseline within-school race and gender gaps, which includes school-by-year FE, controls for teacher characteristics, and controls for observation order and total observations.[23] Whereas the unit of observation in prior tables was 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. Most notably, central office raters give substantially lower scores than principals, assistant principals, or teachers.

In column 3 we add rater fixed effects. If unobserved characteristics of raters are 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 essentially no change in the race or gender gaps between columns 2 and 3.

[23] 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. 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.

In column 4, 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. With only two racial/ethnic groups (and equivalently for gender), we cannot identify separate matching effects for Black and white teachers. We might expect, however, that the magnitude of the race-match effect varies by the racial composition of the school. Appendix Table A9 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. While the estimated coefficient is positive in each group, it is largest in schools with 30–100% Black teachers, suggesting that the match effect is particularly salient in racially diverse schools.[24]

The second within-school mechanism we consider is race and gender 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.[25] For student assignment to be a mechanism for race and gender biases, however, there must be within-school differences in student composition by teacher race and gender.

[24] Even beyond 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. Comparing column 4 to 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. Given that this race-match effect is small, however, the changes in the estimated gaps are marginal (0.02 SD).

[25] More precisely, it rules out true differences in teacher quality present at time of assignment, though teacher performance may still be affected by the students they are assigned.

In Appendix Table A10, we find that Black teachers are systematically assigned more disadvantaged students. Relative to their white colleagues in the same school, Black teachers are assigned more Black students, more FRPL-eligible students, fewer gifted students, more students receiving special education services, more students with prior-year suspensions, and students with lower baseline attendance and achievement scores. Differences between men and women are smaller in magnitude except for student gender, where men tend to have more male students.

Given these patterns, we anticipate that controlling for student assignments would reduce the Black–white observation score gap, with little to no change in the male–female gap. This prediction is borne out in Table 6. A comparison of columns 1 and 2 shows that controlling for assigned student characteristics reduces the estimated Black–white gaps from -0.17 SD to -0.14 (0–25% Black colleagues) and -0.09 to -0.05 (25–50% Black colleagues), while the male–female gap remains essentially unchanged. Columns 3–5 repeat this exercise with a restricted sample of teachers in grades where students have prior test scores, though we find that further controlling for test scores does not appreciably change the estimated gaps.

As a final within-school mechanism we consider the possibility that within-school observation score gaps reflect race and gender differences in teachers' subject and grade assignments. As an example, men in our sample are more likely to teach career and technical education, which could explain part of the gender gap if such teachers tend to receive lower observation scores. Despite some within-school differences in subject and grade assignments (particularly by gender), Appendix Table A11 shows that controlling for these assignments does not appreciably change the race gap and only slightly lowers the gender gap.[26]

[26] There are gender sorting patterns that work in competing directions here. Within a school, women are more likely to teach self-contained classrooms as well as English/Language Arts. Self-contained teachers tend to receive lower observation scores, while English/Language Arts teacher score the highest, on average.

Discussion and Conclusions

As in prior research in other settings (e.g., Campbell & Ronfeldt, 2018; Drake et al., 2019), 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 0.30 SD lower than women, on average.

To assess whether these gaps in part my reflect bias—systematic deviations between observation ratings and actual instructional effectiveness—we estimate models of ratings that account for alternative measures of instructional effectiveness: teacher qualifications, test score value-added, attendance value-added, and, in a robustness check with a small subsample of teachers, scores from student surveys. In all cases, residual gaps favoring white teachers and female teachers persist. Moreover, they are present in all four domains of the rubric and when comparisons are limited to teachers working in the same school in the same year. We interpret this evidence as consistent with, though not definitive proof of, bias in teachers' observation ratings with respect to race and gender. Our preferred estimate of these apparent biases (from Table 2, column 8) are 0.14 SD for race and 0.23 SD for gender.

As one means to see the policy relevance of these estimates, we conducted a simple exercise to assess how much these residual differences could matter for a teacher's summative level of effectiveness (LOE) rating, the final 1–5 score that may trigger personnel action, such as dismissal or tenure denial, at a local school district's discretion. LOE combines observation, TVAAS, and achievement ratings. For four years of data, we have access to the continuous value that produces the LOE rating, so we consider how many Black and male teachers would have moved up a full LOE point if we credited them with the residual gap of 0.14 or 0.23 SD, respectively, in the observation component of the rating, holding the others constant. We find that 1,358 Black teachers (6%) in those two years would have gotten higher LOE with this increase. Impacts would be even larger for

low-scoring teachers; 18% of Black teachers receiving an LOE of 1 and 13% of Black teachers at LOE 2 would have received a higher score, potentially removing the threat of personnel action. For male teachers, 9% overall would have moved up an LOE point, including 33% of male teachers at LOE 1 and 23% at LOE 2.

Our investigations of the potential drivers of the race gap and the gender gap yield different results. For the Black–white observation score gap, school context appears especially important. Sorting of Black and white teachers across schools with different characteristics partially explains the 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. Moreover, we find that average gaps mask substantial heterogeneity by the composition of the school's faculty; 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. School context, however, is not the only driver. Within schools, unlike in (Steinberg & Sartain, 2020), teachers are rated higher when observed by a same-race observer. Teachers also are rated higher when they are assigned to teach fewer historically marginalized students, which again advantages white teachers, who are assigned fewer students of color, low-income students, special education students, and students with histories of lower achievement, lower attendance, and disciplinary action, relative to their Black colleagues in the same school. Even accounting for assigned student characteristics, however, a gap between Black and white teachers remains in most schools. This finding marks a departure from Campbell and Ronfeldt (2018), whose analysis of MET data concludes that accounting for characteristics of a teacher's students makes the Black–white ratings gap statistically indistinguishable from zero, and Steinberg and Sartain (2020), who find a similar pattern in Chicago once school fixed effects are included.

We summarize our findings empirically in Table 7. Specifically, we estimate models similar to those estimated in earlier tables on a common sample of teachers with non-missing covariates to examine how much of the descriptive gap remains as we add

successive sets of covariates. The sample (at the teacher-by-year-by-observation level) includes teachers for whom we can estimate test score value-added, with results for a broader set of teachers in Appendix Table B9. 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.[27] Columns 2 and 3 show evidence of bias in this sample, which column 4 suggests is partially explained by school sorting. Columns 5 and 6 show evidence that rater and assigned student characteristics further partially explain gaps. Accounting for all characteristics at once (column 8), we see that we can explain roughly half of the Black–white gap in schools with fewer than 50% Black colleagues. Black teachers in schools with a majority of Black teachers are predicted to score somewhat higher than their white colleagues, though only about 10% of Tennessee teachers work in such schools.

We have comparatively less success in explaining the gender gap, which seems not to be driven much by school context or other factors that inform the racial gap. Comparing columns 3 and 4 in Table 7, for example, shows that adding school-by-year fixed effects does essentially nothing to the gender gap not explained by test score VA or other teacher characteristics. Neither rater characteristics (include gender matching) nor assigned student characteristics change the estimate. The one partial factor we identify is subject/grade assignment; men are somewhat more likely to teach subject/grade combinations where average ratings are lower. Comparing the full model (column 8) to column 3, we find that including these measures of potential drivers of gender bias reduces the estimated gap by approximately 15%, though it remains substantial.

Our findings have several implications. Foremost, our results raise concerns that Black and male teachers may be disadvantaged in scoring of classroom observations relative to their white and female colleagues who perform similarly on other measures of instructional effectiveness. These disadvantages appear to be driven in part by factors not

[27] All models include controls for observation order and total number of observations.

under teachers' control, such as characteristics of their teaching assignment (e.g., subject, who the students are) within the school. Inaccurate ratings erode the quality of feedback teachers receive about their instruction, which impacts the usefulness of this information for refining their practice. Moreover, in high-stakes contexts like the one in this study, ratings can affect personnel decisions such as contract renewal and compensation. Addressing bias is important to ensure that teachers are treated fairly in evaluation and other personnel processes irrespective of race or gender. Given higher propensities of lower-rated teachers to exit the profession (Drake et al., 2019), combating observation bias may be especially salient in the context of growing calls to diversify a teaching workforce that is overwhelmingly white and female (e.g., Meckler & Rabinowitz, 2019).

How can these apparent biases we document be addressed? Some prior work has suggested that, to offset biases against teachers of some student subgroups, observation scores could 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.

What districts or policymakers should do instead depends on what unobserved factors produce the bias estimates we show. Examples might include rater bias or bias in the rubric itself. We do uncover suggestive evidence that rater bias may be present; raters give higher ratings to teachers of the same race, for example, and in results not shown, we also found evidence that principals rate teachers they hired themselves more favorably than otherwise-similar teachers hired by another principal (results available upon request). If

raters, regardless of their own characteristics, hold implicit (or explicit) biases that favor white and female teachers, districts or states might implement bias training, or they might 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. For instance, the current TEAM rubric does not explicitly consider culturally relevant pedagogy or other approaches to meeting the needs of students from different identity groups. A rubric with attention to such strategies may produce less biased scores.

Future work in this area might delve further into sources of the gaps we document, particularly using data that can illuminate processes that are unobservable in our administrative records, to better guide these recommendations. Research might also explore these patterns in other contexts. 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 from other state or district systems. Such work might help unpack some of the dissimilarities between our results and those in earlier studies (e.g., Steinberg & Sartain, 2020). Also, 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 would be expected.

References

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.

Chi, O. L. (2021). A Classroom Observer Like Me: The Effect of Demographic Congruence Between Teachers and Raters on Observation Scores. *Wheelock Educational Policy Center Working Paper*, 1–62.

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

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.

Gershenson, S. (2016). Linking Teacher Quality, Student Attendance, and Student Achievement. *Education Finance and Policy*, *11*(2), 125–149.

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.

Jann, B. (2008). The Blinder-Oaxaca decomposition for linear regression models. *Stata Journal*, *8*(4), 453–479.

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.

Liu, J., & Loeb, S. (2019). Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School. *Journal of Human Resources*.

Meckler, L., & Rabinowitz, K. (2019, Dec). America's schools are more diverse than ever. but the teachers are still mostly white. *Washington Post*. Retrieved from https://www.washingtonpost.com/graphics/2019/local/education/teacher-diversity/

Milanowski, A. (2017). Lower Performance Evaluation Practice Ratings for Teachers of Disadvantaged Students. *AERA Open*, *3*(1), 1–16.

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence.

Journal of Business and Economic Statistics, *37*(2), 187–204.

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 Measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317.

Steinberg, M. P., & Sartain, L. (2020). What Explains the Race Gap in Teacher Performance Ratings? Evidence From Chicago Public Schools. *Educational Evaluation & Policy Analysis*, *XX*(X), 1–23.

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.

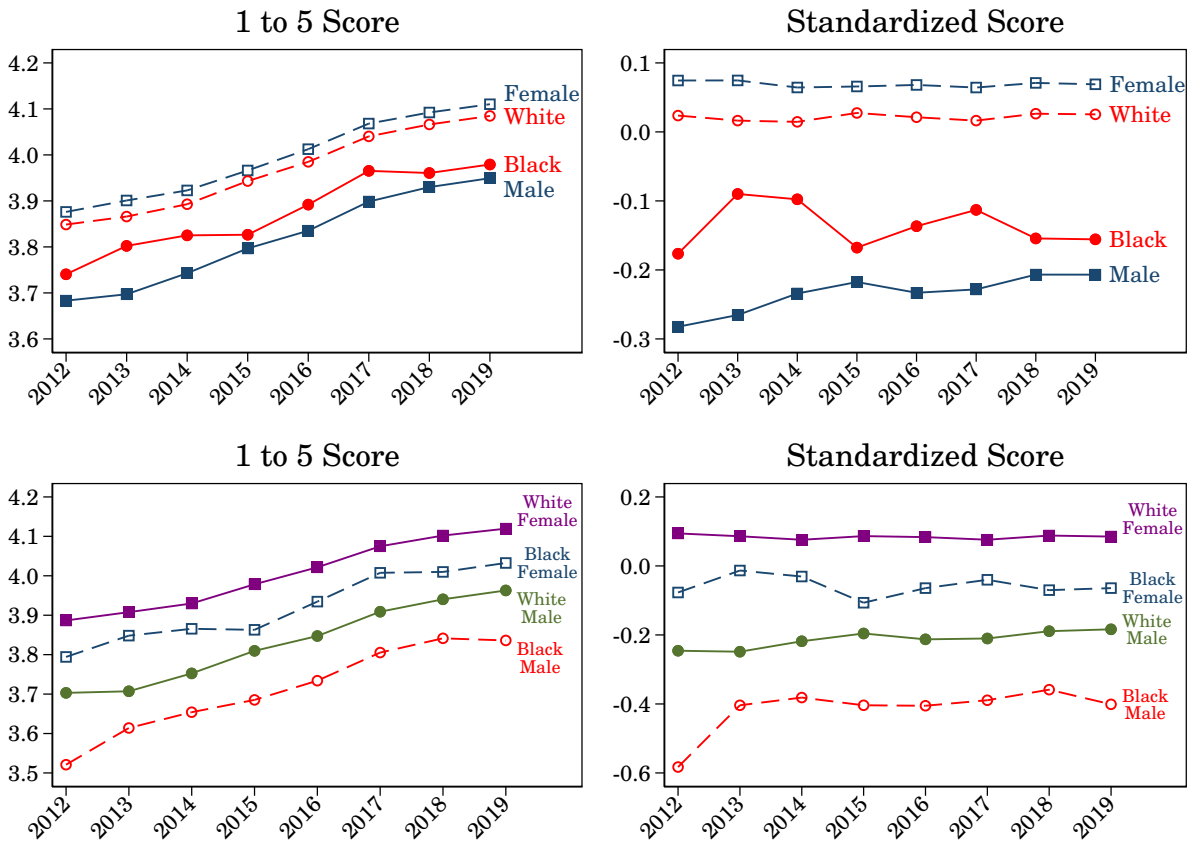

Figure 1. Average Observation Scores by Race and Gender

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.

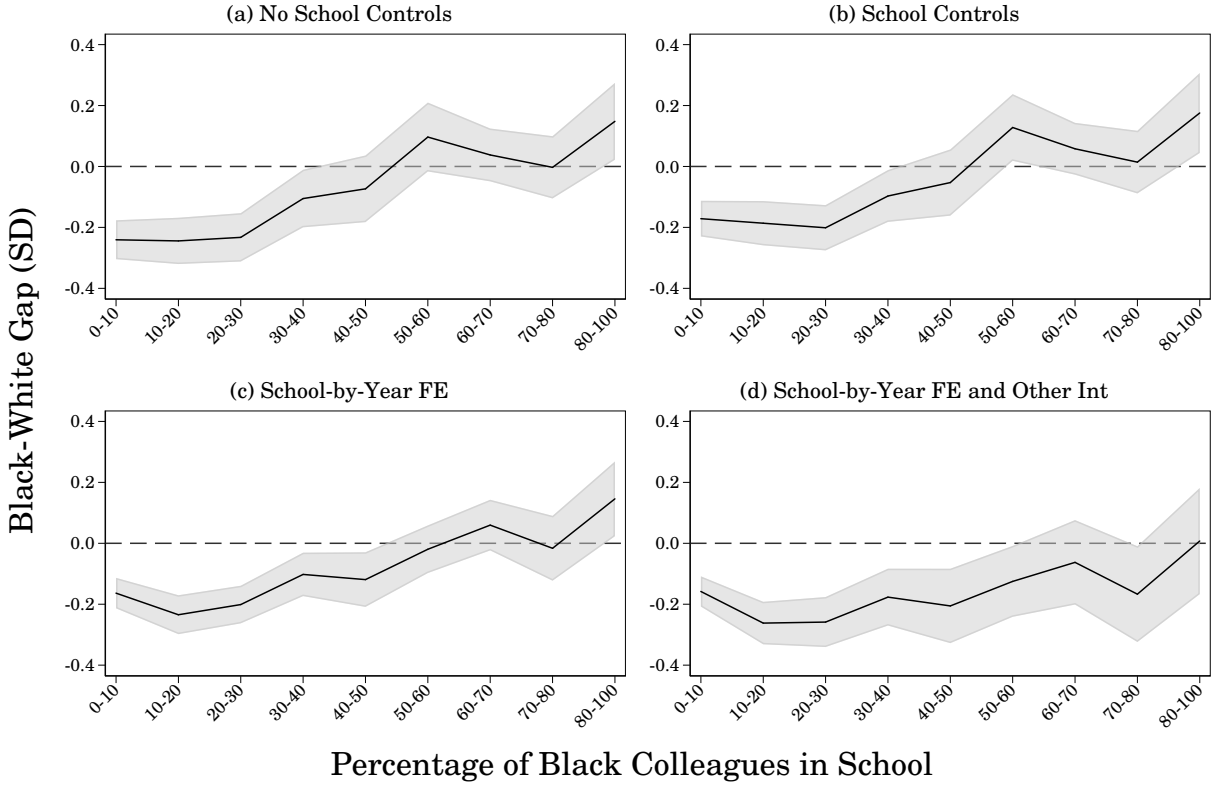

Figure 2. Black–White Gaps in Observation Scores by Teacher Racial Composition in School

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 controls for teacher experience, educational attainment, and test score value-added. 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.

Table 1
Gaps in Standardized Observation Scores by Subgroups

	(A) Race + Gender		(B) Race × Gender		
	Black	Male	Black Female	White Male	Black Male
School Level					
Elementary	-0.122	-0.234	-0.119	-0.231	-0.364
Middle	-0.176	-0.291	-0.160	-0.281	-0.499
High	-0.209	-0.283	-0.183	-0.275	-0.522
School Locale					
Urban	*-0.020*	-0.245	*0.009*	-0.209	-0.328
Suburban	-0.122	-0.309	-0.102	-0.303	-0.505
Town/Rural	-0.365	-0.332	-0.355	-0.330	-0.722
Observation Rubric					
TEAM	-0.391	-0.292	-0.383	-0.289	-0.707
COACH	-0.266	-0.381	-0.225	-0.370	-0.752
TEM	-0.210	-0.306	-0.174	-0.268	-0.482
TIGER	-0.239	-0.291	-0.264	-0.293	-0.464
Rubric Domain (TEAM)					
Instruction	-0.394	-0.263	-0.389	-0.261	-0.669
Environment	-0.352	-0.273	-0.341	-0.270	-0.655
Planning	-0.392	-0.317	-0.396	-0.317	-0.694
Professionalism	-0.343	-0.290	-0.344	-0.290	-0.629
Subject Taught					
Math	-0.176	-0.390	-0.150	-0.379	-0.627
ELA	-0.194	-0.304	-0.169	-0.284	-0.682
Science	-0.198	-0.303	-0.193	-0.299	-0.510
Social Studies	-0.314	-0.295	-0.234	-0.273	-0.753
Self-Contained	-0.155	-0.394	-0.151	-0.386	-0.587
Foreign Language	-0.219	-0.257	-0.229	-0.258	-0.441
Career/Tech Ed	-0.202	-0.364	-0.158	-0.354	-0.620
Arts/Music	-0.196	-0.109	-0.112	-0.090	-0.390
Health/P.E.	*-0.013*	-0.266	-0.096	-0.282	-0.235
Years of Experience					
0–1 Years	-0.215	-0.181	-0.228	-0.187	-0.365
2–4 Years	-0.161	-0.219	-0.157	-0.217	-0.393
5–20 Years	-0.129	-0.300	-0.115	-0.292	-0.479
21+ Years	-0.247	-0.386	-0.247	-0.386	-0.633

Notes: Each combination of row and panel (A and B) shows results from a separate regression model, where the row defines the subsample. 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 2018–19. Italicized estimates are not statistically significant at the 90% level.

Table 2
Do Observation Score Gaps Reflect Differences in Teacher Effectiveness?

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black Teacher	-0.152***	-0.182***	-0.149***	-0.191***	-0.122***	-0.196***	-0.087**	-0.136***
	(0.023)	(0.022)	(0.025)	(0.023)	(0.032)	(0.030)	(0.034)	(0.031)
Male Teacher	-0.301***	-0.286***	-0.322***	-0.262***	-0.299***	-0.275***	-0.308***	-0.231***
	(0.014)	(0.014)	(0.014)	(0.012)	(0.022)	(0.022)	(0.026)	(0.023)
Teacher Characteristics								
MA Degree		0.106***		0.083***		0.111***		0.073***
		(0.007)		(0.008)		(0.011)		(0.013)
MA+ Degree		0.176***		0.138***		0.173***		0.135***
		(0.016)		(0.017)		(0.026)		(0.030)
EdS Degree		0.198***		0.182***		0.218***		0.154***
		(0.017)		(0.018)		(0.026)		(0.028)
PhD Degree		0.245***		0.204***		0.261***		0.214***
		(0.032)		(0.036)		(0.063)		(0.067)
Exp 0–4 years		-0.473***		-0.398***		-0.416***		-0.360***
		(0.011)		(0.012)		(0.018)		(0.020)
Exp 5–14 years		-0.031***		-0.024***		-0.019		-0.011
		(0.007)		(0.009)		(0.012)		(0.015)
Exp 25–39 years		0.026**		0.026**		0.002		0.013
		(0.010)		(0.013)		(0.018)		(0.023)
Exp 40+ years		-0.007		-0.029		-0.065		-0.120
		(0.035)		(0.048)		(0.068)		(0.105)
Drift-Adjusted Value-Added								
Test Score				0.262***				0.318***
				(0.005)				(0.008)
Attendance						0.077***		0.041***
						(0.009)		(0.009)
N	463076	463076	237501	237501	157000	157000	78239	78239
R^2	0.017	0.073	0.020	0.143	0.006	0.055	0.008	0.138
Δ in Black Tch estimate (p)		0.000		0.000		0.000		0.000
Δ in Male Tch estimate (p)		0.151		0.000		0.000		0.000

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. Columns 1, 3, 5, and 7 show the baseline gap estimates for the sample corresponding to the next column(s). Sample sizes become progressively smaller because we cannot estimate value-added for all teachers. The bottom two rows show p-values from tests of equality for the Black Teacher and Male Teacher coefficients in adjacent columns using seemingly unrelated regression with standard errors robust to clustering by school.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3
Observation Score Gaps and School Context

	(1)	(2)	(3)	(4)	(5)
Black Teacher	-0.191***	0.026	-0.132***	-0.318***	-0.238***
	(0.023)	(0.022)	(0.015)	(0.040)	(0.029)
Male Teacher	-0.262***	-0.244***	-0.262***	-0.242***	-0.261***
	(0.012)	(0.009)	(0.008)	(0.009)	(0.008)
Black Teacher x Prop. Black Students				0.629***	0.217***
				(0.073)	(0.050)
School Characteristics					
Enrollment (100s)		0.007**		0.008**	
		(0.003)		(0.003)	
Prop. Black Students		-0.320***		-0.463***	
		(0.054)		(0.059)	
Prop. Hispanic Students		-0.426***		-0.266**	
		(0.120)		(0.121)	
Prop. Gifted Students		1.290***		1.487***	
		(0.430)		(0.440)	
Prop. SPED Students		-0.349*		-0.297	
		(0.205)		(0.205)	
Prop. FRPL Students		-0.161***		-0.180***	
		(0.042)		(0.042)	
Middle School		-0.163***		-0.160***	
		(0.028)		(0.027)	
High School		-0.137***		-0.134***	
		(0.033)		(0.032)	
Other School		-0.036		-0.036	
		(0.065)		(0.065)	
Urban School		0.012		0.017	
		(0.039)		(0.039)	
Town School		0.089**		0.094**	
		(0.038)		(0.038)	
Suburban School		0.002		0.005	
		(0.034)		(0.034)	
Teacher Characteristics	✓	✓	✓	✓	✓
School-by-Year FE			✓		✓
N	237501	237501	237501	237501	237501
R^2	0.143	0.162	0.427	0.165	0.427

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. Teacher characteristics include experience level and educational attainment.
* p < 0.1, ** p < 0.05, *** p < 0.01.

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

	(1)	(2)	(3)	(4)	(5)
Black Teacher	-0.236***	-0.223***	-0.232***	-0.231***	
	(0.028)	(0.030)	(0.022)	(0.034)	
Black Teacher (0–25% Black Colleagues)					-0.196***
					(0.021)
Black Teacher (25–50% Black Colleagues)					-0.121***
					(0.029)
Black Teacher (50–100% Black Colleagues)					0.026
					(0.028)
Interactions					
Black Tch. x Prop. Black Students	0.216***			-0.028	
	(0.050)			(0.077)	
Black Tch. x Prop. Black Colleagues		0.345***		0.228**	
		(0.059)		(0.104)	
Black Tch. x Prop. Black Admin			0.220***	0.130**	
			(0.037)	(0.052)	
Teacher Characteristics	✓	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓	✓
N	238213	238213	238213	238213	238213
R^2	0.427	0.427	0.427	0.427	0.427

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. Teacher characteristics include experience level and educational attainment.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5
Do Rater Characteristics Explain Observation Score Gaps?

	(1)	(2)	(3)	(4)
Black Teacher (0–25% Black Colleagues)	-0.153***	-0.153***	-0.149***	-0.134***
	(0.020)	(0.020)	(0.020)	(0.020)
Black Teacher (25–50% Black Colleagues)	-0.111***	-0.110***	-0.111***	-0.117***
	(0.026)	(0.026)	(0.026)	(0.026)
Black Teacher (50–100% Black Colleagues)	0.036	0.037	0.036	0.017
	(0.033)	(0.033)	(0.034)	(0.035)
Male Teacher	-0.210***	-0.211***	-0.211***	-0.211***
	(0.008)	(0.008)	(0.007)	(0.007)
Rater Characteristics				
Black		-0.017		
		(0.016)		
Male		0.020**		
		(0.009)		
Ed.S. Degree		0.003	0.007	0.007
		(0.011)	(0.023)	(0.023)
Ph.D. Degree		-0.026*	0.010	0.011
		(0.015)	(0.027)	(0.027)
Assistant Principal		0.013	0.029*	0.029*
		(0.010)	(0.017)	(0.017)
Teacher		0.032*	-0.053	-0.053
		(0.019)	(0.041)	(0.041)
Central Office		-0.168***	-0.121***	-0.121***
		(0.027)	(0.038)	(0.038)
3–5 Years Admin Exp.		-0.003	-0.020**	-0.020**
		(0.009)	(0.010)	(0.010)
6–9 Years Admin Exp.		0.010	-0.036***	-0.036***
		(0.011)	(0.014)	(0.014)
10+ Years Admin Exp.		0.030**	-0.028*	-0.028*
		(0.012)	(0.017)	(0.017)
Race Match w/ Teacher				0.032***
				(0.011)
Gender Match w/ Teacher				0.007
				(0.005)
Observation Order				
Second	0.171***	0.171***	0.170***	0.170***
	(0.005)	(0.005)	(0.004)	(0.004)
Third	0.399***	0.398***	0.393***	0.393***
	(0.009)	(0.009)	(0.008)	(0.008)
Fourth	0.492***	0.491***	0.486***	0.486***
	(0.012)	(0.012)	(0.011)	(0.011)
Fifth or more	0.643***	0.639***	0.631***	0.631***
	(0.017)	(0.017)	(0.016)	(0.016)
Total Observations				
Three	-0.498***	-0.497***	-0.492***	-0.492***
	(0.008)	(0.008)	(0.008)	(0.008)
Four	-0.825***	-0.823***	-0.814***	-0.814***
	(0.013)	(0.013)	(0.013)	(0.013)
Five or more	-0.975***	-0.972***	-0.964***	-0.963***
	(0.014)	(0.014)	(0.013)	(0.013)
Teacher Characteristics	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓
Rater FE			✓	✓
N	485548	485548	485307	485307
R^2	0.383	0.384	0.424	0.424

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. School-level clustered standard errors shown in parentheses. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. Teacher characteristics include experience level and educational attainment.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6
Do Within-School Student Assignments Explain Teacher Observation Score Gaps?

			Achievement Sample		
	(1)	(2)	(3)	(4)	(5)
Black Teacher (0–25% Black Colleagues)	-0.172***	-0.142***	-0.178***	-0.147***	-0.146***
	(0.022)	(0.022)	(0.023)	(0.023)	(0.023)
Black Teacher (25–50% Black Colleagues)	-0.092***	-0.049	-0.127***	-0.081**	-0.082***
	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
Black Teacher (50–100% Black Colleagues)	0.066**	0.092***	0.031	0.062*	0.061*
	(0.030)	(0.029)	(0.032)	(0.032)	(0.032)
Male Teacher	-0.251***	-0.249***	-0.255***	-0.252***	-0.251***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Assigned Student Characteristics					
Prop. Female Students		0.230***		0.245***	0.208***
		(0.028)		(0.033)	(0.033)
Prop. Amer Ind Students		-0.468**		-0.159	-0.147
		(0.209)		(0.263)	(0.262)
Prop. Asian Students		0.318**		0.340**	0.235
		(0.126)		(0.144)	(0.145)
Prop. Black Students		-0.253***		-0.278***	-0.183***
		(0.049)		(0.058)	(0.058)
Prop. Hispanic Students		-0.024		-0.151*	-0.121
		(0.073)		(0.077)	(0.079)
Prop. Pac Isl Students		-0.304		-0.232	-0.286
		(0.292)		(0.417)	(0.419)
Prop. FRPL Students		-0.661***		-0.584***	-0.477***
		(0.030)		(0.037)	(0.038)
Prop. ELL Students		0.136*		0.301***	0.490***
		(0.075)		(0.077)	(0.080)
Prop. Gifted Students		0.662***		0.490***	0.343***
		(0.085)		(0.077)	(0.074)
Prop. SPED Students		0.050**		-0.011	0.140***
		(0.023)		(0.027)	(0.030)
Prop. Prior-year ISS		-0.216***		-0.229***	-0.172***
		(0.050)		(0.053)	(0.053)
Prop. Prior-year OSS		-0.030		-0.121	-0.095
		(0.067)		(0.076)	(0.075)
Prop. Prior-year Expel		-0.016		-0.088	-0.103
		(0.365)		(0.348)	(0.341)
Prop. Prior-year Retain		-0.617***		-0.423**	-0.382**
		(0.118)		(0.193)	(0.191)
Prior-year Absences (std)		-0.111***		-0.142***	-0.100***
		(0.019)		(0.030)	(0.029)
Prior-year Math (std)					0.082***
					(0.013)
Prior-year ELA (std)					0.061***
					(0.015)
Teacher Characteristics	✓	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓	✓
N	205042	205042	174376	174376	174376
R^2	0.452	0.463	0.471	0.481	0.482

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. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. Teacher characteristics include experience level and educational attainment.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7
How Much of Observation Score Gaps Can We Explain?

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black Teacher (0–25% Black Colleagues)	-0.205***	-0.223***	-0.231***	-0.125***	-0.106***	-0.105***	-0.121***	-0.084***
	(0.030)	(0.030)	(0.029)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Black Teacher (25–50% Black Colleagues)	-0.115***	-0.140***	-0.140***	-0.095***	-0.098***	-0.069**	-0.096***	-0.074***
	(0.032)	(0.032)	(0.031)	(0.028)	(0.028)	(0.028)	(0.028)	(0.027)
Black Teacher (50–100% Black Colleagues)	0.085**	0.058	0.036	0.057*	0.035	0.076**	0.066**	0.065**
	(0.037)	(0.037)	(0.038)	(0.034)	(0.034)	(0.034)	(0.033)	(0.033)
Male Teacher	-0.227***	-0.228***	-0.199***	-0.208***	-0.207***	-0.207***	-0.176***	-0.170***
	(0.012)	(0.012)	(0.011)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Teacher Characteristics		✓	✓	✓	✓	✓	✓	✓
Test Score Value-Added			✓	✓	✓	✓	✓	✓
School-by-Year FE				✓	✓	✓	✓	✓
Rater Characteristics					✓			✓
Assigned Student Characteristics						✓		✓
Subject/Grade Assignment							✓	✓
N	420312	420312	420312	420312	418487	420312	420312	418487
R^2	0.176	0.179	0.208	0.396	0.436	0.402	0.404	0.449

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. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. All models include controls for the main effect of colleague race, observation order, and the total number of observations received in that year. Teacher characteristics include experience level and educational attainment. Rater characteristics include educational attainment, job title, admin experience, rater fixed effects, and binary indicators for race and gender match. Columns 5 and 8 differ in sample size due to dropping of singleton observations.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Appendix A

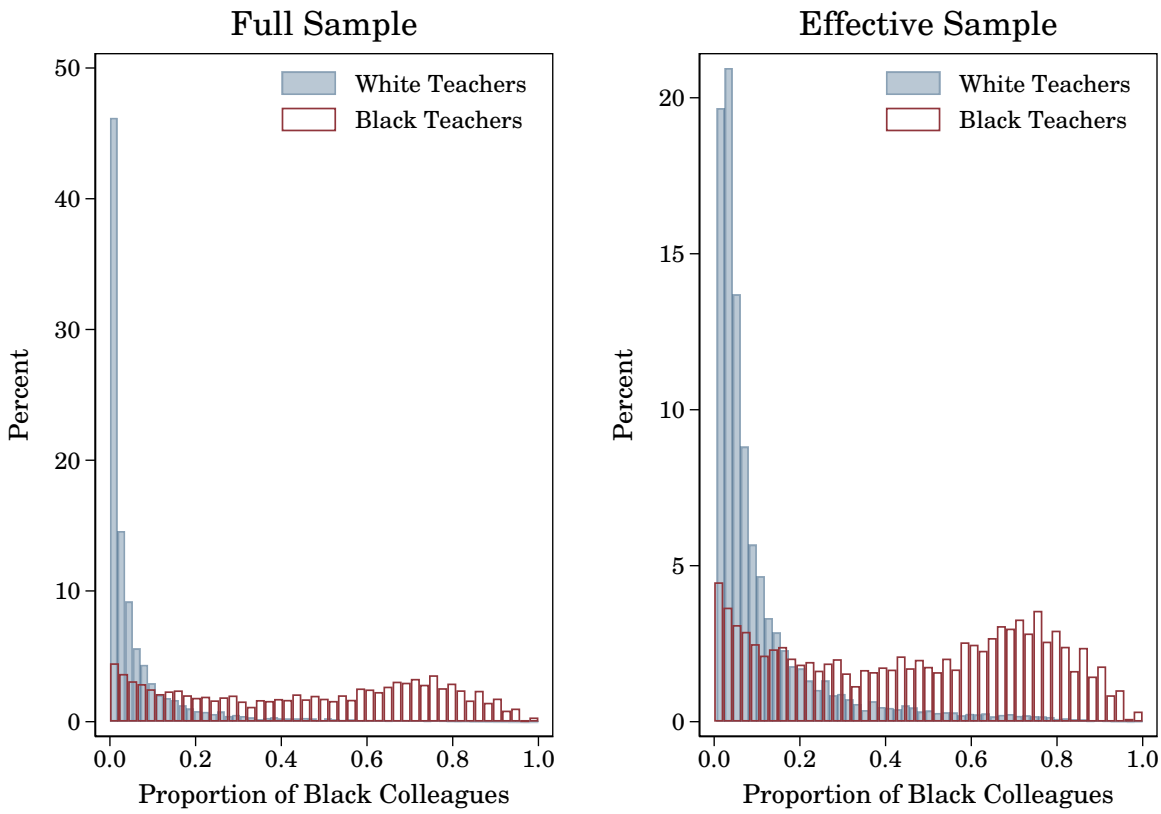

Figure A1. Distributions of Colleague Race

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%, 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 A1
Average Teacher, School, Colleague, and Observation Characteristics by Race and Gender

	All	Black	White	Female	Male
Teacher Characteristics					
Black	0.11			0.11	0.10
White	0.89			0.89	0.90
Female	0.79	0.80	0.79		
Male	0.21	0.20	0.21		
Age	42.6	44.2	42.5	42.6	43.0
Years of Experience	12.1	12.5	12.0	12.2	11.6
MA Degree	0.41	0.37	0.42	0.42	0.38
MA+ Degree	0.06	0.13	0.05	0.06	0.06
EdS Degree	0.07	0.09	0.07	0.07	0.07
PhD Degree	0.01	0.01	0.01	0.01	0.01
Test Score Value-Added	0.02	0.05	0.02	0.06	-0.12
Attendance Value-Added	-0.02	0.25	-0.05	-0.01	-0.05
School Characteristics					
Enrollment (100s)	8.37	8.42	8.37	7.92	10.09
Prop. Black Students	0.22	0.65	0.17	0.22	0.22
Prop. Hispanic Students	0.09	0.12	0.08	0.09	0.08
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	0.52	0.71	0.50	0.53	0.49
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	0.25	0.26	0.19	0.52
Other School	0.04	0.02	0.04	0.03	0.06
Urban School	0.30	0.75	0.25	0.30	0.30
Suburban School	0.20	0.10	0.21	0.20	0.19
Town School	0.17	0.05	0.19	0.17	0.17
Rural School	0.33	0.10	0.36	0.33	0.34
Colleague Characteristics					
Prop. Black Colleagues	0.11	0.46	0.07	0.11	0.11
Prop. Male Colleagues	0.21	0.21	0.21	0.18	0.31
Black Principal	0.16	0.64	0.10	0.16	0.15
Male Principal	0.47	0.43	0.47	0.43	0.59
Prop. Black Administrators	0.18	0.64	0.12	0.18	0.18
Prop. Male Administrators	0.42	0.37	0.42	0.39	0.53
Observation Characteristics					
Average Observation Score (1 to 5)	3.96	3.87	3.97	3.99	3.82
Average Observation Score (SD)	0.00	-0.14	0.02	0.07	-0.23
Total Classroom Observations in Year	3.1	3.1	3.0	3.0	3.1
Total Raters in Year	1.9	1.9	1.9	1.8	1.9
N (Teacher-Year)	475107	51192	423915	375119	99486

Notes: Sample includes all Tennessee teachers with non-missing average observation scores from 2011–12 to 2018–19. Due to the very small number, we also drop non-Black, non-white teachers from the analysis.

Table A2
Do Observation Score Gaps Reflect Differences in Teacher Effectiveness? (w/ Student Surveys)

					Value-Added Sample					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Black Teacher	-0.067	-0.114***	-0.036	-0.097***	-0.039	-0.094*	-0.123***	-0.039	-0.085**	-0.084**
	(0.045)	(0.041)	(0.030)	(0.027)	(0.051)	(0.048)	(0.045)	(0.040)	(0.038)	(0.034)
Male Teacher	-0.182***	-0.104***	-0.137***	-0.084***	-0.317***	-0.255***	-0.207***	-0.295***	-0.249***	-0.207***
	(0.040)	(0.037)	(0.026)	(0.024)	(0.044)	(0.042)	(0.039)	(0.033)	(0.031)	(0.027)
Student Survey Score										
1		-0.976***		-1.030***		-0.976***	-0.814***		-0.993***	-0.825***
		(0.107)		(0.102)		(0.155)	(0.150)		(0.151)	(0.150)
2		-0.512***		-0.543***		-0.632***	-0.544***		-0.629***	-0.547***
		(0.037)		(0.030)		(0.046)	(0.045)		(0.041)	(0.040)
3		-0.251***		-0.225***		-0.269***	-0.224***		-0.242***	-0.206***
		(0.020)		(0.017)		(0.025)	(0.024)		(0.022)	(0.022)
4 (base)										
5		0.222***		0.145***		0.245***	0.209***		0.201***	0.169***
		(0.029)		(0.020)		(0.031)	(0.030)		(0.024)	(0.024)
Drift-Adjusted Test Score VA							0.288***			0.256***
							(0.013)			(0.011)
School-by-Year FE			✓	✓				✓	✓	✓
Teacher Characteristics		✓		✓		✓	✓		✓	✓
N	21216	21216	21216	21216	10908	10908	10908	10908	10908	10908
R^2	0.006	0.110	0.308	0.374	0.015	0.105	0.194	0.346	0.404	0.460

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. Columns 1–4 contain the full sample of teachers with student survey scores. Columns 5–10 restrict to teachers with both student survey scores and test score VA.
* p < 0.1, ** p < 0.05, *** p < 0.01

Table A3
Observation Score Gaps by Rubric Domains (TEAM rubric only)

	Instruction		Environment		Planning		Professionalism	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black Teacher	-0.396***	-0.375***	-0.349***	-0.330***	-0.414***	-0.398***	-0.366***	-0.350***
	(0.028)	(0.027)	(0.027)	(0.026)	(0.026)	(0.025)	(0.028)	(0.028)
Male Teacher	-0.269***	-0.221***	-0.279***	-0.236***	-0.306***	-0.268***	-0.277***	-0.239***
	(0.014)	(0.013)	(0.015)	(0.013)	(0.014)	(0.013)	(0.016)	(0.015)
Teacher Characteristics								
MA Degree	0.115***	0.096***	0.089***	0.072***	0.100***	0.085***	0.118***	0.104***
	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)	(0.008)	(0.010)	(0.009)
MA+ Degree	0.148***	0.127***	0.072***	0.053***	0.133***	0.117***	0.165***	0.149***
	(0.023)	(0.020)	(0.021)	(0.019)	(0.021)	(0.019)	(0.022)	(0.020)
EdS Degree	0.181***	0.185***	0.148***	0.151***	0.174***	0.177***	0.220***	0.223***
	(0.022)	(0.021)	(0.019)	(0.018)	(0.020)	(0.020)	(0.022)	(0.022)
PhD Degree	0.223***	0.230***	0.075*	0.080**	0.213***	0.218***	0.178***	0.183***
	(0.042)	(0.040)	(0.042)	(0.039)	(0.037)	(0.036)	(0.044)	(0.042)
Exp 0–4 years	-0.394***	-0.382***	-0.346***	-0.336***	-0.249***	-0.239***	-0.264***	-0.254***
	(0.014)	(0.014)	(0.014)	(0.013)	(0.014)	(0.013)	(0.015)	(0.014)
Exp 5–14 years	-0.018*	-0.024**	-0.024**	-0.029***	-0.000	-0.005	0.022**	0.017
	(0.011)	(0.010)	(0.010)	(0.009)	(0.010)	(0.010)	(0.011)	(0.011)
Exp 25–39 years	0.026	0.037**	0.015	0.025*	0.013	0.022	-0.010	-0.001
	(0.016)	(0.015)	(0.014)	(0.014)	(0.015)	(0.015)	(0.016)	(0.015)
Exp 40+ years	-0.072	-0.032	-0.045	-0.010	-0.106*	-0.075	-0.123*	-0.092
	(0.059)	(0.055)	(0.052)	(0.051)	(0.057)	(0.055)	(0.064)	(0.063)
Drift-Adjusted Value-Added								
Test Score		0.263***		0.234***		0.205***		0.205***
		(0.006)		(0.005)		(0.006)		(0.006)
N	164304	164304	164304	164304	164304	164304	164304	164304
R^2	0.063	0.144	0.050	0.113	0.045	0.094	0.045	0.093

Notes: In each model, the dependent variable is a teacher's average observation score in the given year for the rubric domain listed in the header. Scores are standardized within year. Models estimated via OLS. School-level clustered standard errors shown in parentheses. The sample includes teachers in districts that used the TEAM rubric between 2013–2017.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A4
Observation Score Gaps by Rubric Domains (TEAM rubric only)

	Instruction		Environment		Planning		Professionalism	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black Teacher	-0.392***	-0.357***	-0.358***	-0.329***	-0.421***	-0.395***	-0.355***	-0.330***
	(0.041)	(0.037)	(0.041)	(0.039)	(0.038)	(0.036)	(0.037)	(0.036)
Male Teacher	-0.243***	-0.180***	-0.263***	-0.211***	-0.286***	-0.239***	-0.223***	-0.179***
	(0.030)	(0.026)	(0.029)	(0.026)	(0.027)	(0.025)	(0.028)	(0.026)
Teacher Characteristics								
MA Degree	0.112***	0.084***	0.077***	0.055***	0.088***	0.067***	0.115***	0.096***
	(0.016)	(0.015)	(0.015)	(0.014)	(0.015)	(0.014)	(0.015)	(0.014)
MA+ Degree	0.205***	0.154***	0.089**	0.047	0.177***	0.138***	0.199***	0.163***
	(0.044)	(0.038)	(0.038)	(0.034)	(0.038)	(0.034)	(0.036)	(0.033)
EdS Degree	0.135***	0.157***	0.089***	0.106***	0.145***	0.162***	0.212***	0.227***
	(0.037)	(0.033)	(0.032)	(0.028)	(0.035)	(0.032)	(0.039)	(0.035)
PhD Degree	0.249***	0.236***	0.119*	0.109*	0.205***	0.196***	0.215***	0.206***
	(0.079)	(0.075)	(0.063)	(0.060)	(0.070)	(0.068)	(0.076)	(0.074)
Exp 0–4 years	-0.343***	-0.334***	-0.264***	-0.257***	-0.199***	-0.193***	-0.208***	-0.202***
	(0.025)	(0.023)	(0.023)	(0.021)	(0.023)	(0.022)	(0.023)	(0.022)
Exp 5–14 years	-0.003	-0.008	-0.001	-0.006	0.006	0.002	0.028	0.024
	(0.019)	(0.017)	(0.017)	(0.016)	(0.017)	(0.017)	(0.017)	(0.017)
Exp 25–39 years	0.062**	0.048*	0.044*	0.033	0.046	0.036	0.010	0.001
	(0.031)	(0.028)	(0.026)	(0.025)	(0.028)	(0.026)	(0.029)	(0.027)
Exp 40+ years	-0.101	-0.026	-0.109	-0.048	0.023	0.078	-0.174	-0.122
	(0.161)	(0.138)	(0.130)	(0.116)	(0.135)	(0.130)	(0.131)	(0.119)
Drift-Adjusted Value-Added								
Test Score		0.332***		0.272***		0.249***		0.233***
		(0.009)		(0.008)		(0.009)		(0.009)
Attendance		0.047***		0.037***		0.037***		0.034***
		(0.010)		(0.009)		(0.010)		(0.010)
N	52847	52847	52847	52847	52847	52847	52847	52847
R^2	0.043	0.148	0.031	0.106	0.030	0.089	0.030	0.086

Notes: In each model, the dependent variable is a teacher's average observation score in the given year for the rubric domain listed in the header. Scores are standardized within year. Models estimated via OLS. School-level clustered standard errors shown in parentheses. The sample includes teachers in districts that used the TEAM rubric between 2013–2017.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A5
Do Observation Score Gaps Reflect Differences in Teacher Effectiveness? (School-by-Year FE)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black Teacher	-0.085***	-0.140***	-0.088***	-0.130***	-0.057**	-0.120***	-0.037	-0.080***
	(0.012)	(0.012)	(0.017)	(0.015)	(0.022)	(0.022)	(0.031)	(0.027)
Male Teacher	-0.271***	-0.255***	-0.301***	-0.263***	-0.277***	-0.264***	-0.291***	-0.237***
	(0.007)	(0.007)	(0.009)	(0.008)	(0.019)	(0.019)	(0.024)	(0.020)
Teacher Characteristics								
MA Degree		0.118***		0.104***		0.126***		0.092***
		(0.005)		(0.006)		(0.009)		(0.012)
MA+ Degree		0.118***		0.114***		0.123***		0.098***
		(0.010)		(0.013)		(0.021)		(0.027)
EdS Degree		0.180***		0.147***		0.202***		0.133***
		(0.009)		(0.011)		(0.018)		(0.022)
PhD Degree		0.249***		0.219***		0.242***		0.170***
		(0.024)		(0.029)		(0.052)		(0.058)
Exp 0–4 years		-0.380***		-0.302***		-0.294***		-0.231***
		(0.007)		(0.009)		(0.013)		(0.017)
Exp 5–14 years		-0.007		0.002		0.023**		0.035***
		(0.006)		(0.007)		(0.010)		(0.013)
Exp 25–39 years		0.025***		0.019*		0.007		0.005
		(0.008)		(0.011)		(0.015)		(0.021)
Exp 40+ years		-0.003		-0.032		-0.061		-0.144
		(0.031)		(0.045)		(0.058)		(0.107)
Drift-Adjusted Value-Added								
Test Score				0.247***				0.309***
				(0.004)				(0.007)
Attendance						0.053***		0.033***
						(0.005)		(0.006)
School-by-Year FE	✓	✓	✓	✓	✓	✓	✓	✓
N	468764	468764	240117	240117	158735	158735	79131	79131
R^2	0.329	0.365	0.344	0.427	0.358	0.384	0.398	0.480

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. Columns 1, 3, 5, and 7 show the baseline gap estimates for the sample corresponding to the next column(s). Sample sizes become progressively smaller because we cannot estimate value-added for all teachers.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A6

Do Observation Score Gaps Reflect Differences in Teacher Effectiveness? (Alternative Value-Added Measure)

	(1)	(2)	(3)	(4)	(5)	(6)
Black Teacher	-0.149***	-0.147***	-0.178***	-0.090***	-0.123***	-0.120***
	(0.029)	(0.026)	(0.026)	(0.021)	(0.018)	(0.017)
Male Teacher	-0.330***	-0.261***	-0.241***	-0.302***	-0.254***	-0.242***
	(0.014)	(0.012)	(0.012)	(0.011)	(0.010)	(0.009)
Teacher Characteristics						
MA Degree		0.103***	0.095***		0.103***	0.097***
		(0.009)	(0.009)		(0.008)	(0.008)
MA+ Degree		0.168***	0.155***		0.109***	0.102***
		(0.021)	(0.020)		(0.016)	(0.015)
EdS Degree		0.178***	0.184***		0.145***	0.134***
		(0.021)	(0.021)		(0.014)	(0.014)
PhD Degree		0.251***	0.244***		0.255***	0.247***
		(0.042)	(0.041)		(0.035)	(0.034)
Exp 0–4 years		-0.360***	-0.366***		-0.271***	-0.276***
		(0.015)	(0.015)		(0.011)	(0.011)
Exp 5–14 years		-0.039***	-0.042***		-0.012	-0.015*
		(0.011)	(0.011)		(0.009)	(0.009)
Exp 25–39 years		0.029*	0.032**		0.032**	0.032**
		(0.016)	(0.016)		(0.014)	(0.013)
Exp 40+ years		-0.009	-0.000		0.010	0.009
		(0.062)	(0.059)		(0.056)	(0.054)
Value-Added						
TVAAS Index		0.331***	0.229***		0.319***	0.226***
		(0.006)	(0.007)		(0.006)	(0.006)
Drift-Adjusted Test Score VA			0.171***			0.167***
			(0.006)			(0.005)
School-by-Year FE				✓	✓	✓
N	109184	109184	109184	109184	109184	109184
R^2	0.021	0.183	0.207	0.371	0.486	0.504

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. Columns 1 and 4 show the baseline gap estimates for the sample corresponding to the next column(s).

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A7

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

	Subgroup			
	Black (1)	White (2)	Female (3)	Male (4)
Test Score Value-Added	0.267***	0.257***	0.254***	0.251***
	(0.010)	(0.006)	(0.005)	(0.008)
Teacher Characteristics				
MA Degree	0.070***	0.091***	0.080***	0.094***
	(0.026)	(0.008)	(0.008)	(0.016)
MA+ Degree	0.085***	0.176***	0.151***	0.151***
	(0.033)	(0.019)	(0.018)	(0.033)
EdS Degree	0.203***	0.166***	0.169***	0.221***
	(0.038)	(0.019)	(0.018)	(0.031)
PhD Degree	0.113	0.244***	0.217***	0.238***
	(0.093)	(0.036)	(0.038)	(0.065)
Exp 0–4 years	-0.393***	-0.357***	-0.392***	-0.270***
	(0.029)	(0.012)	(0.013)	(0.020)
Exp 5–14 years	-0.018	-0.017*	-0.026***	0.023
	(0.024)	(0.009)	(0.009)	(0.017)
Exp 25–39 years	-0.066	0.049***	0.041***	-0.044
	(0.042)	(0.013)	(0.014)	(0.029)
Exp 40+ years	-0.101	-0.017	-0.019	-0.062
	(0.127)	(0.051)	(0.057)	(0.083)
School Characteristics				
Enrollment (100s)	0.014**	0.007**	0.007**	0.008**
	(0.005)	(0.003)	(0.003)	(0.004)
Prop. Black Students	0.441***	-0.447***	-0.268***	-0.419***
	(0.092)	(0.059)	(0.049)	(0.070)
Prop. Hispanic Students	0.475**	-0.380***	-0.378***	-0.599***
	(0.205)	(0.120)	(0.118)	(0.178)
Prop. Gifted Students	2.975***	1.153**	1.452***	0.615
	(0.487)	(0.456)	(0.424)	(0.586)
Prop. SPED Students	0.260	-0.320	-0.386**	-0.195
	(0.284)	(0.218)	(0.184)	(0.353)
Prop. FRPL Students	-0.384***	-0.188***	-0.154***	-0.203***
	(0.093)	(0.042)	(0.041)	(0.062)
Middle School	-0.244***	-0.181***	-0.157***	-0.194***
	(0.053)	(0.029)	(0.027)	(0.038)
High School	-0.292***	-0.195***	-0.133***	-0.166***
	(0.057)	(0.034)	(0.033)	(0.042)
Other School	-0.068	-0.089	-0.009	-0.116
	(0.111)	(0.065)	(0.062)	(0.077)
Urban School	0.018	0.012	0.001	0.050
	(0.063)	(0.040)	(0.037)	(0.056)
Town School	0.107	0.100***	0.099***	0.049
	(0.096)	(0.038)	(0.037)	(0.056)
Suburban School	-0.049	0.008	0.005	-0.005
	(0.080)	(0.035)	(0.033)	(0.051)
N	26576	210931	187060	50441
R^2	0.148	0.158	0.149	0.146

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.

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A8
Observation Score Gaps and School Context (Effective sample of schools with at least one Black teacher)

	(1)	(2)	(3)	(4)	(5)
Black Teacher	-0.136***	0.008	-0.135***	-0.288***	-0.239***
	(0.023)	(0.022)	(0.015)	(0.039)	(0.029)
Male Teacher	-0.275***	-0.267***	-0.282***	-0.263***	-0.280***
	(0.016)	(0.012)	(0.010)	(0.012)	(0.010)
Black Teacher x Prop. Black Students				0.547***	0.213***
				(0.072)	(0.050)
School Characteristics					
Enrollment (100s)		0.010***		0.010***	
		(0.004)		(0.004)	
Prop. Black Students		-0.150**		-0.295***	
		(0.065)		(0.072)	
Prop. Hispanic Students		-0.243*		-0.094	
		(0.145)		(0.145)	
Prop. Gifted Students		1.698***		1.864***	
		(0.448)		(0.460)	
Prop. SPED Students		-0.237		-0.170	
		(0.256)		(0.255)	
Prop. FRPL Students		-0.254***		-0.271***	
		(0.059)		(0.058)	
Middle School		-0.170***		-0.168***	
		(0.035)		(0.035)	
High School		-0.137***		-0.134***	
		(0.042)		(0.042)	
Other School		-0.055		-0.059	
		(0.112)		(0.112)	
Urban School		0.020		0.024	
		(0.046)		(0.047)	
Town School		0.118**		0.128**	
		(0.054)		(0.054)	
Suburban School		-0.022		-0.015	
		(0.051)		(0.051)	
Teacher Characteristics	✓	✓	✓	✓	✓
School-by-Year FE			✓		✓
N	141867	141867	141867	141867	141867
R^2	0.144	0.168	0.423	0.171	0.423

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. Teacher characteristics include experience level and educational attainment.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A9
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	-0.061*	-0.065**	-0.121***	-0.122***	-0.034	-0.042
	(0.033)	(0.031)	(0.028)	(0.027)	(0.028)	(0.028)
Male Teacher	-0.156***	-0.156***	-0.221***	-0.216***	-0.251***	-0.249***
	(0.009)	(0.009)	(0.020)	(0.020)	(0.029)	(0.028)
Black Rater	0.054		-0.008		-0.016	
	(0.034)		(0.028)		(0.025)	
Male Rater	0.018		0.020		0.093***	
	(0.012)		(0.024)		(0.032)	
Race Match w/ Teacher	0.024	0.015	0.027	0.020	0.060***	0.068***
	(0.026)	(0.023)	(0.019)	(0.018)	(0.019)	(0.022)
Gender Match w/ Teacher	0.007	0.008	-0.001	0.005	0.003	-0.008
	(0.006)	(0.006)	(0.014)	(0.013)	(0.022)	(0.022)
Teacher Characteristics	✓	✓	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓	✓	✓
Rater FE		✓		✓		✓
N	335203	335023	56159	56071	27321	27211
R^2	0.405	0.443	0.393	0.438	0.393	0.446

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. School-level clustered standard errors shown in parentheses. Teacher characteristics include experience level and educational attainment. Rater characteristics include job title, educational attainment, and admin experience.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A10
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	-0.010***	0.022***	0.018***	-0.008***	0.023***	0.011***	0.013***	0.022***	-0.069***	-0.073***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.001)	(0.003)	(0.009)	(0.010)
Male Teacher	-0.011***	0.001*	-0.005***	0.000	-0.010***	0.008***	0.004***	-0.002	-0.000	0.002
	(0.001)	(0.001)	(0.001)	(0.000)	(0.002)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)
Teacher Characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Within-School SD	0.097	0.053	0.089	0.022	0.126	0.043	0.037	0.179	0.294	0.287
N	205189	205180	205189	205189	205189	205080	205080	205072	174432	174410
R^2	0.178	0.952	0.875	0.436	0.142	0.690	0.748	0.589	0.539	0.553

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. 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. Teacher characteristics include experience level and educational attainment.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A11
Do Subject and Grade Assignments Explain Observation Score Gaps?

	(1)	(2)	(3)	(4)
Black Teacher (0–25% Black Colleagues)	-0.142***	-0.137***	-0.141***	-0.137***
	(0.022)	(0.021)	(0.022)	(0.021)
Black Teacher (25–50% Black Colleagues)	-0.049	-0.050	-0.056*	-0.054*
	(0.032)	(0.031)	(0.031)	(0.031)
Black Teacher (50–100% Black Colleagues)	0.092***	0.092***	0.087***	0.091***
	(0.029)	(0.028)	(0.029)	(0.028)
Male Teacher	-0.250***	-0.208***	-0.259***	-0.213***
	(0.008)	(0.008)	(0.008)	(0.008)
Subject Taught (Proportion)				
Math		-0.039***		-0.038***
		(0.010)		(0.010)
ELA		ref.		ref.
Science		-0.173***		-0.174***
		(0.011)		(0.011)
Social Studies		-0.243***		-0.246***
		(0.013)		(0.013)
Self-Contained		-0.282***		-0.107***
		(0.015)		(0.016)
Foreign Language		-0.266***		-0.264***
		(0.052)		(0.052)
Career/Tech Ed		-0.068***		-0.050**
		(0.021)		(0.021)
Arts/Music		-0.007		0.032
		(0.039)		(0.039)
Health/P.E.		-0.100***		-0.072**
		(0.028)		(0.029)
Grade Taught (Proportion)				
Pre-K			-0.216*	-0.176
			(0.119)	(0.120)
Kindergarten			-0.215***	-0.180***
			(0.048)	(0.050)
Grade 1			-0.241***	-0.207***
			(0.045)	(0.046)
Grade 2			-0.297***	-0.264***
			(0.043)	(0.043)
Grade 3			-0.059	-0.025
			(0.041)	(0.041)
Grade 4			0.015	0.044
			(0.040)	(0.040)
Grade 5			0.072*	0.097**
			(0.040)	(0.040)
Grade 6			0.016	0.036
			(0.037)	(0.037)
Grade 7			0.020	0.036
			(0.036)	(0.036)
Grade 8			0.117***	0.133***
			(0.036)	(0.036)
Grade 9			ref.	ref.
Grade 10			-0.029	-0.027
			(0.020)	(0.020)
Grade 11			0.010	0.051**
			(0.020)	(0.020)
Grade 12			-0.015	-0.029
			(0.022)	(0.022)
Teacher Characteristics	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓
Assigned Student Controls	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓
N	205006	205006	205006	205006
R^2	0.463	0.470	0.468	0.473

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. 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. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. Teacher characteristics include experience level and educational attainment.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Appendix B
Results for Tables 3–7 Using Alternative Samples

Table B1
Observation Score Gaps and School Context (Table 3 with Full Sample)

	(1)	(2)	(3)	(4)	(5)
Black Teacher	-0.182***	-0.004	-0.141***	-0.373***	-0.296***
	(0.022)	(0.020)	(0.012)	(0.034)	(0.021)
Male Teacher	-0.286***	-0.253***	-0.255***	-0.250***	-0.254***
	(0.014)	(0.008)	(0.007)	(0.008)	(0.007)
Black Teacher x Prop. Black Students				0.670***	0.310***
				(0.065)	(0.037)
School Characteristics					
Enrollment (100s)		0.011***		0.012***	
		(0.003)		(0.003)	
Prop. Black Students		-0.196***		-0.337***	
		(0.055)		(0.059)	
Prop. Hispanic Students		-0.497***		-0.349***	
		(0.120)		(0.120)	
Prop. Gifted Students		1.809***		2.010***	
		(0.429)		(0.438)	
Prop. SPED Students		-0.122		-0.080	
		(0.197)		(0.196)	
Prop. FRPL Students		-0.333***		-0.350***	
		(0.041)		(0.041)	
Middle School		-0.185***		-0.185***	
		(0.029)		(0.029)	
High School		-0.196***		-0.197***	
		(0.035)		(0.034)	
Other School		-0.080		-0.084	
		(0.073)		(0.073)	
Urban School		0.014		0.019	
		(0.040)		(0.040)	
Town School		0.060		0.066*	
		(0.038)		(0.038)	
Suburban School		-0.024		-0.021	
		(0.034)		(0.034)	
Teacher Characteristics	✓	✓	✓	✓	✓
School-by-Year FE			✓		✓
N	463076	463076	463076	463076	463076
R^2	0.073	0.100	0.365	0.103	0.365

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. Teacher characteristics include experience level and educational attainment.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table B2
Observation Score Gaps and School Context (Table 3 with Attendance and Test Score VA Sample)

	(1)	(2)	(3)	(4)	(5)
Black Teacher	-0.136***	0.132***	-0.082***	-0.288***	-0.207***
	(0.031)	(0.034)	(0.028)	(0.061)	(0.052)
Male Teacher	-0.231***	-0.204***	-0.236***	-0.203***	-0.236***
	(0.023)	(0.022)	(0.020)	(0.021)	(0.020)
Black Teacher x Prop. Black Students				0.756***	0.260***
				(0.106)	(0.090)
School Characteristics					
Enrollment (100s)		0.007		0.009*	
		(0.005)		(0.005)	
Prop. Black Students		-0.415***		-0.605***	
		(0.071)		(0.080)	
Prop. Hispanic Students		-0.262**		-0.072	
		(0.133)		(0.135)	
Prop. Gifted Students		2.653***		3.048***	
		(0.520)		(0.534)	
Prop. SPED Students		-0.202		-0.087	
		(0.237)		(0.236)	
Prop. FRPL Students		-0.136***		-0.158***	
		(0.051)		(0.050)	
Middle School		-0.137***		-0.133***	
		(0.034)		(0.033)	
High School		-0.155**		-0.164**	
		(0.074)		(0.073)	
Other School		0.119		0.112	
		(0.078)		(0.076)	
Urban School		-0.011		0.002	
		(0.045)		(0.045)	
Town School		0.097**		0.107**	
		(0.045)		(0.045)	
Suburban School		0.011		0.018	
		(0.039)		(0.039)	
Teacher Characteristics	✓	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓	✓
Attendance Value-Added	✓	✓	✓	✓	✓
School-by-Year FE			✓		✓
N	78239	78239	78239	78239	78239
R^2	0.138	0.163	0.480	0.167	0.480

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. Teacher characteristics include experience level and educational attainment.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table B3

Disentangling Black Students, Black Teachers, and Black Administrators (Table 4 with Full Sample)

	(1)	(2)	(3)	(4)	(5)
Black Teacher	-0.294***	-0.266***	-0.262***	-0.290***	
	(0.021)	(0.021)	(0.018)	(0.024)	
Black Teacher (0–25% Black Colleagues)					-0.230***
					(0.015)
Black Teacher (25–50% Black Colleagues)					-0.108***
					(0.026)
Black Teacher (50–100% Black Colleagues)					0.032
					(0.021)
Interactions					
Black Tch. x Prop. Black Students	0.309***			0.061	
	(0.037)			(0.059)	
Black Tch. x Prop. Black Colleagues		0.432***		0.234***	
		(0.044)		(0.085)	
Black Tch. x Prop. Black Admin			0.262***	0.123***	
			(0.029)	(0.044)	
Teacher Characteristics	✓	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓	✓
N	464544	464544	464544	464544	464544
R^2	0.364	0.365	0.365	0.365	0.365

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. Teacher characteristics include experience level and educational attainment.

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table B4

Disentangling Black Students, Black Teachers, and Black Administrators (Table 4 with Attendance and Test Score VA Sample)

	(1)	(2)	(3)	(4)	(5)
Black Teacher	-0.194***	-0.125*	-0.198***	-0.108	
	(0.056)	(0.075)	(0.046)	(0.080)	
Black Teacher (0–25% Black Colleagues)					-0.174***
					(0.041)
Black Teacher (25–50% Black Colleagues)					-0.072
					(0.060)
Black Teacher (50–100% Black Colleagues)					0.072
					(0.064)
Interactions					
Black Tch. x Prop. Black Students	0.189*			-0.158	
	(0.097)			(0.152)	
Black Tch. x Prop. Black Colleagues		0.465***		0.507**	
		(0.123)		(0.222)	
Black Tch. x Prop. Black Admin			0.214***	0.103	
			(0.073)	(0.089)	
Teacher Characteristics	✓	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓	✓
Attendance Value-Added	✓	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓	✓
N	59638	59638	59638	59638	59638
R^2	0.480	0.481	0.480	0.481	0.481

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. Teacher characteristics include experience level and educational attainment.

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table B5

Do Rater Characteristics Explain Observation Score Gaps? (Table 5 with Full Sample)

	(1)	(2)	(3)	(4)
Black Teacher (0–25% Black Colleagues)	-0.176***	-0.176***	-0.172***	-0.153***
	(0.014)	(0.014)	(0.014)	(0.014)
Black Teacher (25–50% Black Colleagues)	-0.094***	-0.093***	-0.093***	-0.101***
	(0.021)	(0.021)	(0.021)	(0.021)
Black Teacher (50–100% Black Colleagues)	0.021	0.022	0.022	-0.003
	(0.025)	(0.025)	(0.025)	(0.026)
Male Teacher	-0.204***	-0.205***	-0.206***	-0.206***
	(0.006)	(0.006)	(0.006)	(0.006)
Rater Characteristics				
Black		-0.017		
		(0.014)		
Male		0.022**		
		(0.009)		
Ed.S. Degree		-0.000	-0.015	-0.015
		(0.010)	(0.018)	(0.018)
Ph.D. Degree		-0.020	-0.019	-0.019
		(0.014)	(0.024)	(0.024)
Assistant Principal		0.020**	0.026*	0.026*
		(0.008)	(0.014)	(0.013)
Teacher		0.025	-0.024	-0.024
		(0.017)	(0.033)	(0.033)
Central Office		-0.092***	-0.131***	-0.131***
		(0.021)	(0.029)	(0.029)
3–5 Years Admin Exp.		-0.004	-0.022***	-0.022***
		(0.007)	(0.008)	(0.008)
6–9 Years Admin Exp.		0.002	-0.037***	-0.037***
		(0.010)	(0.011)	(0.011)
10+ Years Admin Exp.		0.028***	-0.038***	-0.038***
		(0.011)	(0.013)	(0.013)
Race Match w/ Teacher				0.039***
				(0.009)
Gender Match w/ Teacher				0.001
				(0.004)
Observation Order				
Second	0.164***	0.164***	0.164***	0.164***
	(0.004)	(0.004)	(0.004)	(0.004)
Third	0.372***	0.373***	0.371***	0.371***
	(0.008)	(0.008)	(0.008)	(0.008)
Fourth	0.472***	0.473***	0.470***	0.470***
	(0.010)	(0.010)	(0.010)	(0.010)
Fifth or more	0.581***	0.581***	0.574***	0.574***
	(0.014)	(0.014)	(0.014)	(0.014)
Total Observations				
Three	-0.562***	-0.562***	-0.557***	-0.557***
	(0.008)	(0.008)	(0.007)	(0.007)
Four	-0.914***	-0.914***	-0.905***	-0.905***
	(0.012)	(0.012)	(0.012)	(0.012)
Five or more	-1.064***	-1.064***	-1.054***	-1.054***
	(0.012)	(0.012)	(0.011)	(0.011)
Teacher Characteristics	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓
Rater FE			✓	✓
N	938968	938968	938803	938803
R^2	0.348	0.348	0.384	0.384

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. School-level clustered standard errors shown in parentheses. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. Teacher characteristics include experience level and educational attainment.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table B6

Do Rater Characteristics Explain Observation Score Gaps? (Table 5 with Test Score and Attendance VA Sample)

	(1)	(2)	(3)	(4)
Black Teacher (0–25% Black Colleagues)	-0.136***	-0.136***	-0.132***	-0.108***
	(0.035)	(0.035)	(0.034)	(0.034)
Black Teacher (25–50% Black Colleagues)	-0.089*	-0.088	-0.089	-0.103*
	(0.054)	(0.054)	(0.055)	(0.053)
Black Teacher (50–100% Black Colleagues)	0.095*	0.096*	0.113*	0.081
	(0.057)	(0.057)	(0.061)	(0.063)
Male Teacher	-0.179***	-0.179***	-0.179***	-0.169***
	(0.019)	(0.019)	(0.019)	(0.019)
Rater Characteristics				
Black		-0.031		
		(0.021)		
Male		0.012		
		(0.014)		
Ed.S. Degree		0.003	-0.039	-0.040
		(0.016)	(0.041)	(0.041)
Ph.D. Degree		-0.038*	0.020	0.020
		(0.020)	(0.059)	(0.059)
Assistant Principal		0.015	-0.024	-0.024
		(0.011)	(0.032)	(0.032)
Teacher		0.020	0.102	0.099
		(0.027)	(0.085)	(0.085)
Central Office		-0.070**	-0.083	-0.083
		(0.035)	(0.099)	(0.099)
3–5 Years Admin Exp.		0.012	-0.011	-0.010
		(0.013)	(0.016)	(0.016)
6–9 Years Admin Exp.		0.023	-0.007	-0.007
		(0.016)	(0.025)	(0.025)
10+ Years Admin Exp.		0.030	-0.030	-0.030
		(0.018)	(0.029)	(0.029)
Race Match w/ Teacher				0.049**
				(0.022)
Gender Match w/ Teacher				0.026*
				(0.015)
Observation Order				
Second	0.176***	0.176***	0.176***	0.176***
	(0.006)	(0.006)	(0.006)	(0.006)
Third	0.423***	0.423***	0.422***	0.422***
	(0.011)	(0.011)	(0.011)	(0.011)
Fourth	0.504***	0.504***	0.508***	0.508***
	(0.015)	(0.015)	(0.015)	(0.015)
Fifth or more	0.685***	0.683***	0.683***	0.683***
	(0.020)	(0.020)	(0.020)	(0.020)
Total Observations				
Three	-0.515***	-0.515***	-0.510***	-0.510***
	(0.013)	(0.013)	(0.013)	(0.013)
Four	-0.865***	-0.864***	-0.854***	-0.854***
	(0.024)	(0.024)	(0.023)	(0.023)
Five or more	-0.996***	-0.994***	-0.988***	-0.988***
	(0.020)	(0.020)	(0.020)	(0.020)
Teacher Characteristics	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓
Attendance Value-Added	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓
Rater FE			✓	✓
N	150610	150610	150084	150084
R^2	0.422	0.422	0.461	0.461

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. School-level clustered standard errors shown in parentheses. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. Teacher characteristics include experience level and educational attainment.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B7

Do Within-School Student Assignments Explain Teacher Observation Score Gaps? (Table 6 with Full Sample)

			Achievement Sample		
	(1)	(2)	(3)	(4)	(5)
Black Teacher (0–25% Black Colleagues)	-0.215***	-0.179***	-0.199***	-0.159***	-0.159***
	(0.017)	(0.017)	(0.020)	(0.019)	(0.019)
Black Teacher (25–50% Black Colleagues)	-0.085***	-0.042	-0.093***	-0.044	-0.045
	(0.030)	(0.029)	(0.033)	(0.031)	(0.031)
Black Teacher (50–100% Black Colleagues)	0.042**	0.063***	0.041	0.069***	0.067***
	(0.021)	(0.021)	(0.026)	(0.025)	(0.025)
Male Teacher	-0.240***	-0.231***	-0.252***	-0.243***	-0.243***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Assigned Student Characteristics					
Prop. Female Students		0.175***		0.146***	0.116***
		(0.019)		(0.024)	(0.024)
Prop. Amer Ind Students		-0.357**		-0.080	-0.079
		(0.145)		(0.248)	(0.248)
Prop. Asian Students		0.123		0.443***	0.320**
		(0.077)		(0.130)	(0.125)
Prop. Black Students		-0.372***		-0.456***	-0.352***
		(0.039)		(0.051)	(0.051)
Prop. Hispanic Students		-0.036		-0.132**	-0.104
		(0.043)		(0.062)	(0.063)
Prop. Pac Isl Students		-0.250		-0.654*	-0.720**
		(0.181)		(0.359)	(0.362)
Prop. FRPL Students		-0.781***		-0.651***	-0.542***
		(0.022)		(0.036)	(0.036)
Prop. ELL Students		-0.117***		-0.126**	0.109*
		(0.040)		(0.054)	(0.058)
Prop. Gifted Students		0.911***		0.572***	0.389***
		(0.098)		(0.088)	(0.084)
Prop. SPED Students		0.099***		0.010	0.200***
		(0.016)		(0.025)	(0.027)
Prop. Prior-year ISS		-0.333***		-0.290***	-0.217***
		(0.045)		(0.050)	(0.049)
Prop. Prior-year OSS		-0.166***		-0.210***	-0.185**
		(0.058)		(0.076)	(0.074)
Prop. Prior-year Expel		-0.537*		-0.626*	-0.637*
		(0.316)		(0.345)	(0.338)
Prop. Prior-year Retain		-0.140**		-0.621***	-0.589***
		(0.057)		(0.150)	(0.148)
Prior-year Absences (std)		-0.101***		-0.247***	-0.199***
		(0.009)		(0.024)	(0.023)
Prior-year Math (std)					0.076***
					(0.011)
Prior-year ELA (std)					0.092***
					(0.012)
Teacher Characteristics	✓	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓	✓
N	384149	384149	261284	261284	261284
R^2	0.381	0.395	0.397	0.411	0.413

Notes: 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. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. Teacher characteristics include experience level and educational attainment.

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table B8

Do Within-School Student Assignments Explain Teacher Observation Score Gaps? (Table 6 with Attendance and Test Score VA Sample)

			Achievement Sample		
	(1)	(2)	(3)	(4)	(5)
Black Teacher (0–25% Black Colleagues)	-0.150***	-0.135***	-0.154***	-0.143***	-0.142***
	(0.038)	(0.038)	(0.041)	(0.041)	(0.041)
Black Teacher (25–50% Black Colleagues)	-0.009	0.021	-0.081	-0.053	-0.052
	(0.055)	(0.055)	(0.056)	(0.056)	(0.056)
Black Teacher (50–100% Black Colleagues)	0.139**	0.136**	0.093	0.097	0.095
	(0.055)	(0.054)	(0.073)	(0.074)	(0.074)
Male Teacher	-0.219***	-0.226***	-0.205***	-0.204***	-0.203***
	(0.020)	(0.020)	(0.022)	(0.022)	(0.022)
Assigned Student Characteristics					
Prop. Female Students		0.195***		0.183***	0.166**
		(0.047)		(0.064)	(0.065)
Prop. Amer Ind Students		-0.650**		-0.195	-0.201
		(0.265)		(0.435)	(0.436)
Prop. Asian Students		-0.021		-0.255	-0.268
		(0.174)		(0.192)	(0.193)
Prop. Black Students		-0.206***		-0.174**	-0.125
		(0.072)		(0.087)	(0.089)
Prop. Hispanic Students		0.023		-0.105	-0.097
		(0.114)		(0.109)	(0.110)
Prop. Pac Isl Students		-0.225		0.237	0.210
		(0.400)		(0.701)	(0.703)
Prop. FRPL Students		-0.627***		-0.404***	-0.350***
		(0.044)		(0.063)	(0.065)
Prop. ELL Students		-0.006		0.234*	0.355***
		(0.123)		(0.120)	(0.123)
Prop. Gifted Students		0.957***		0.552***	0.429**
		(0.212)		(0.193)	(0.191)
Prop. SPED Students		0.300***		0.051	0.136**
		(0.041)		(0.060)	(0.061)
Prop. Prior-year ISS		0.051		-0.285*	-0.260*
		(0.128)		(0.150)	(0.151)
Prop. Prior-year OSS		0.286**		-0.079	-0.042
		(0.119)		(0.167)	(0.166)
Prop. Prior-year Expel		-0.061		-0.960	-0.965
		(0.923)		(1.113)	(1.107)
Prop. Prior-year Retain		-0.660***		-0.033	0.107
		(0.155)		(0.340)	(0.342)
Prior-year Absences (std)		-0.090***		-0.081**	-0.061
		(0.025)		(0.039)	(0.039)
Prior-year Math (std)					0.052*
					(0.027)
Prior-year ELA (std)					0.046*
					(0.027)
Teacher Characteristics	✓	✓	✓	✓	✓
Test Score Value-Added	✓	✓	✓	✓	✓
Attendance Value-Added	✓	✓	✓	✓	✓
School-by-Year FE	✓	✓	✓	✓	✓
N	71330	71330	44118	44118	44118
R^2	0.504	0.512	0.584	0.588	0.588

Notes: 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. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. Teacher characteristics include experience level and educational attainment.

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table B9
How Much of Observation Score Gaps Can We Explain? (Table 7 with Full Sample)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Black Teacher (0–25% Black Colleagues)	-0.224***	-0.243***	-0.157***	-0.133***	-0.134***	-0.154***	-0.109***
	(0.023)	(0.023)	(0.015)	(0.016)	(0.015)	(0.015)	(0.015)
Black Teacher (25–50% Black Colleagues)	-0.112***	-0.140***	-0.078***	-0.087***	-0.054**	-0.080***	-0.066***
	(0.029)	(0.029)	(0.024)	(0.023)	(0.024)	(0.024)	(0.023)
Black Teacher (50–100% Black Colleagues)	0.049	0.021	0.025	-0.000	0.036	0.027	0.017
	(0.030)	(0.030)	(0.025)	(0.026)	(0.025)	(0.025)	(0.025)
Male Teacher	-0.210***	-0.209***	-0.201***	-0.203***	-0.196***	-0.194***	-0.183***
	(0.012)	(0.011)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Teacher Characteristics		✓	✓	✓	✓	✓	✓
School-by-Year FE			✓	✓	✓	✓	✓
Rater Characteristics				✓			✓
Assigned Student Characteristics					✓		✓
Subject/Grade Assignment						✓	✓
N	783198	783198	783198	779861	783198	783198	779861
R^2	0.173	0.177	0.359	0.395	0.366	0.364	0.406

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. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. All models include controls for the main effect of colleague race, observation order, and the total number of observations received in that year. Teacher characteristics include experience level and educational attainment. Rater characteristics include educational attainment, job title, admin experience, rater fixed effects, and binary indicators for race and gender match. Columns 4 and 7 differ in sample size due to dropping of singleton observations.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table B10

How Much of Observation Score Gaps Can We Explain? (Table 7 with Test Score and Attendance VA Sample)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black Teacher (0–25% Black Colleagues)	-0.209***	-0.226***	-0.224***	-0.107***	-0.083**	-0.097***	-0.106***	-0.074**
	(0.040)	(0.040)	(0.038)	(0.035)	(0.036)	(0.035)	(0.035)	(0.036)
Black Teacher (25–50% Black Colleagues)	-0.057	-0.081	-0.097*	-0.064	-0.071	-0.047	-0.075	-0.056
	(0.058)	(0.059)	(0.055)	(0.054)	(0.053)	(0.054)	(0.054)	(0.053)
Black Teacher (50–100% Black Colleagues)	0.190***	0.165***	0.152***	0.112*	0.099	0.115*	0.114*	0.106
	(0.060)	(0.061)	(0.057)	(0.066)	(0.072)	(0.065)	(0.063)	(0.068)
Male Teacher	-0.191***	-0.195***	-0.161***	-0.170***	-0.158***	-0.175***	-0.186***	-0.172***
	(0.023)	(0.023)	(0.022)	(0.020)	(0.020)	(0.019)	(0.020)	(0.019)
Teacher Characteristics		✓	✓	✓	✓	✓	✓	✓
Test Score and Attendance Value-Added			✓	✓	✓	✓	✓	✓
School-by-Year FE				✓	✓	✓	✓	✓
Rater Characteristics					✓			✓
Assigned Student Characteristics						✓		✓
Subject/Grade Assignment							✓	✓
N	136693	136693	136693	136693	135575	136693	136693	135575
R^2	0.171	0.174	0.208	0.433	0.470	0.438	0.446	0.484

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. For the Black–white gap, we report the estimated marginal effect in schools with 0–25%, 25–50%, and 50–100% Black colleagues. All models include controls for the main effect of colleague race, observation order, and the total number of observations received in that year. Teacher characteristics include experience level and educational attainment. Rater characteristics include educational attainment, job title, admin experience, rater fixed effects, and binary indicators for race and gender match. Columns 5 and 8 differ in sample size due to dropping of singleton observations.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Appendix C
Bounding Exercise

The results in Table 2 show the Black–white and male–female gaps in observation scores with and without controls for determinants or proxies of teacher effectiveness. We find that substantial gaps remain even conditional on these controls, which include teacher education, years of experience, and value-added to student test score and attendance. As we note, treating these adjusted gaps as evidence of bias requires an assumption that there are no unobserved determinants of instructional effectiveness that are correlated with race or gender. In this section, we conduct a bounding exercise proposed by Oster (2019) to check the robustness of our results to unobserved confounders.

As described in Oster (2019), a common way to assess robustness to the threat of unobservables is to assume that observables and unobservables are similarly correlated (i.e., in the same direction) with the treatment (in this context, teacher race or gender). Based on the change in the estimated treatment effect when controlling for observables along with the change in the R-squared, we can obtain a bias-adjusted treatment effect.

Consider the following model:

$$Y = \beta X + W_1 + W_2 + \epsilon \tag{6}$$

where Y is the observation score, X is a binary indicator for Black or male teacher, W_1 are the observable determinants/proxies of teacher instructional effectiveness (educational attainment, years of experience, value-added), and W_2 is the index of unobservable determinants/proxies of teacher instructional effectiveness. In this setup, we assume that W_1 and W_2 behave similarly–for instance, if controlling for W_1 attenuates the estimate of β, then controlling for W_2 will lead to further attenuation. To obtain the bias-adjusted treatment effect, we must set two parameters: δ and R_{max}. δ is the coefficient of proportionality: $\delta\frac{\sigma_{1X}}{\sigma_1^2} = \frac{\sigma_{2X}}{\sigma_2^2}$, where $\sigma_{iX} = cov(W_i, X)$ and $\sigma_i^2 = var(W_i)$. Intuitively, δ represents the the extent to which unobservables (W_2) are more or less related to race/gender than observables (W_1). Oster (2019) suggests that an appropriate upper bound is $\delta = 1$, whereby observables and unobservables are equally important in predicting race/gender. Our bounding exercise considers values of δ from 0.25 to 1.0.

The second parameter to consider is R_{max}, which is the (theoretical) R^2 from a regression of Y on X, W_1, and W_2. That is, R_{max} is the proportion of total variation in observation scores that is attributable to teachers' instructional effectiveness. In considering appropriate bounds for R_{max}, we first note that observation scores almost certainly contain substantial measurement error, and that some of the variation in observation scores is driven by factors other than teacher effectiveness, such as differences in school context. Thus, the upper bound for R_{max} should be well below 1. One way to obtain a reasonable upper bound is to regress current-year observation scores on prior-year observation scores, based on the notion that teacher effectiveness is largely fixed across years. In our data, this simple regression yields an R^2 of 0.5, which we take as our upper bound for R_{max}. In comparison, this is well above corresponding empirical estimates using test score VA, where

the R^2 ranges from 0.03 to 0.41 (Koedel et al., 2015).[28] Nonetheless, using the year-to-year correlation in observation scores still likely overstates R_{max}, since most teachers remain in the same school in adjacent years. For teachers remaining in the same school, prior-year obervation score explain 53% of the variation in current-year observation scores, compared to 28% for teachers who switch schools. Another plausible R_{max} comes from the year-to-year correlation in observation scores that are first residualized on a vector of school fixed effects, which produces an R^2 of 0.39. Given this range of estimates, we conduct our bounding exercise for values of R_{max} between 0.3 and 0.5.

We implement this bounding exercise using the "psacalc" package in Stata. For the parameters of δ and R_{max} outlined above, we report the bias-adjusted treatment effect for Black teacher and male teacher using the sample that has both test score and attendance VA. We also show results when including school-by-year fixed effects as nuisance parameters (this uses the within-school-year R^2 as the amount of variation explained by observables). While we include the same observables as shown in Table 2, we control for the complete set of indicators for each year of experience (instead of the buckets that include multiple years) and control for cubic functions of the VA measures. This has a negligible effect on the magnitude of the gap estimates but increases the baseline R^2 from 0.138 to 0.157. Tables C1 and C2 show the results for the race and gender gap, respectively. Because the Black–white gap increases in magnitude when controlling for observables, it will further increase unless unobservables work in the opposite direction. As an illustrative example, if R_{max} is 0.4 then δ would have to be -2.8 to completely erase the race gap. For the gender gap, the bias-adjusted estimates remain negative except for the upper bounds of δ and R_{max}. For our preferred R_{max} of 0.4, δ would need to be 1.6 to completely attenuate the male–female gap. At R_{max} of 0.5, δ would need to be 1.1.

Table C1
Bias-Adjusted Estimate for Race Gap (β_{Black})

				With School-by-Year FE		
R_{max} =	0.3	0.4	0.5	0.3	0.4	0.5
δ =						
0.25	-0.128	-0.135	-0.141	-0.098	-0.107	-0.115
0.5	-0.137	-0.150	-0.163	-0.110	-0.126	-0.143
0.75	-0.146	-0.166	-0.187	-0.121	-0.146	-0.172
1.0	-0.156	-0.183	-0.213	-0.132	-0.166	-0.201

Notes: Results shown are bias-adjusted treatment effects calculated according to the approach in Oster (2019). R_{max} is the R-squared from a hypothetical regression of observation scores on observed and unobserved determinants of instructional effectiveness. δ is the coefficient of proportionality, where $\delta > 1$ means that unobservables explain more of the variance in group membership (i.e., race or gender) than observables.

[28] Specifically, Koedel et al. (2015) document that the year-to-year correlation is estimated teacher value-added have produced estimates that change from 0.18 to 0.64.

Table C2
Bias-Adjusted Estimate for Gender Gap (β_{Male})

				With School-by-Year FE		
$R_{max} =$	0.3	0.4	0.5	0.3	0.4	0.5
$\delta =$						
0.25	-0.205	-0.191	-0.177	-0.216	-0.206	-0.196
0.5	-0.185	-0.156	-0.128	-0.203	-0.182	-0.162
0.75	-0.165	-0.122	-0.078	-0.189	-0.159	-0.129
1.0	-0.144	-0.086	-0.028	-0.176	-0.135	-0.094

Notes: Results shown are bias-adjusted treatment effects calculated according to the approach in Oster (2019). R_{max} is the R-squared from a hypothetical regression of observation scores on observed and unobserved determinants of instructional effectiveness. δ is the coefficient of proportionality, where $\delta > 1$ means that unobservables explain more of the variance in group membership (i.e., race or gender) than observables.