FIRST IMPRESSIONS MATTER: EVIDENCE FROM ELEMENTARY-SCHOOL TEACHERS*

MARCOS A. RANGEL
Duke University‡

YING SHI
Syracuse University§

Abstract

We study the empirical relevance of first impressions in the context of education. We find that teachers who begin their careers in classrooms with large White-Black incoming score differentials carry negative views into evaluations of future cohorts of Black students relative to their White classmates. Our evidence is based on novel data on blind evaluations and non-blind public school teacher assessments of fourth and fifth graders in North Carolina. Teachers' perceptions are particularly sensitive to relatively low-performing Black students in early classrooms, but not to relatively high-performing Black students. Since teacher expectations can shape grading patterns and sorting into academic tracks as well as students' own beliefs and behaviors, these findings suggest an important link between specific novice teachers' initial experiences and the persistence of racial gaps in educational attainment and achievement.

JEL: I24, J15

*Data Availability Statement: The data used in this article cannot be distributed by authors but can be obtained from the North Carolina Education Research Data Center (`https://childandfamilypolicy.duke.edu/north-carolina-education-research-data/`). Replication programs will be provided in online repository (DOI to be added). Disclosure Statement: Rangel and Shi both have nothing to disclose. Research conducted under Duke University's IRB Approved Protocol # 2018-0596 (Jun-14-2018). All errors are those of the authors.

‡Rangel: Sanford School of Public Policy, Duke University; 262 Rubenstein Hall, Box 90312, Durham, NC 27708; Email: marcos.rangel@duke.edu; Phone: (919) 613-7340.

§Shi: Department of Public Administration and International Affairs, Syracuse University; 426 Eggers Hall, Syracuse, NY 13244; Email: yshi78@syr.edu

1 Introduction

While researchers have examined racial differentials in the US and international settings, less is known about the role of first impressions in shaping their prevalence and persistence. We study this in the education context by examining the early career experiences of public school teachers. Our focus is on the extent to which first impressions feed into racial differences in subsequent teacher assessments of students' skills. The relevance of teacher evaluations is evident in research showing how teacher expectations matter for students' long-term academic trajectories (Papageorge et al., 2016; Lavy and Sand, 2018). We complement previous contributions which emphasize the role of exposure to particular racial groups (Asch, 1946; Lang, 1986; Cornell and Welch, 1996; Rabin and Schrag, 1999; Ambady and Skowronski, 2008; Pettigrew et al., 2011; Devine et al., 2012) by focusing on the *nature* of this contact, as defined by the ability distributions of students entering teachers' initial classrooms. This reasoning borrows insights from scholars in psychology and economics who have underscored how reliance on stereotypes, or over-generalized representations of group characteristics, can promote the rise in biased judgment (Hilton and von Hippel, 1996; Bordalo et al., 2016; Alesina et al., 2018). Since the empirical literature on this topic lags behind, our work aims at filling an important gap in knowledge.

Our analyses indicate that the distribution of academic abilities by racial group among students entering a teacher's first classroom is a salient element in shaping her beliefs and ultimately influences the way she evaluates other cohorts of students who she interacts with between the second and fourth years of her career.[1] Our analysis is made possible by unique matched student-teacher administrative data from the North Carolina Education Research Data Center (NCERDC). For the years we study, NCERDC contains subjective teacher assessments as well as blind-scored standardized tests covering the same underlying skillsets. The objective test measure provides a reference point for diagnosing whether teachers are differentially evaluating White vs. Black students. Our estimates yield significant within-classroom White-Black gaps even after accounting for dif-

[1] Data availability precludes the extension of our analysis to longer career trajectories. We discuss this in detail below.

ferences in blindly-scored test performance and student demographic attributes. Combining both math and reading in our preferred specification, we find that the average Black student is evaluated at 0.060 points below her comparable White peers on a 1-4 scale, corresponding to 0.07 standard deviations. An alternative interpretation is that the average Black student is 2.5 percentage points less likely to be deemed proficient than an observationally equivalent White classmate (equivalent to 0.05 standard deviations or a 4% reduction relative to average Black proficiency rates). These magnitudes are comparable to previous findings in the grading discrimination literature.[2] We also note that, at least conceptually, one would not expect to find racial differentials that are substantially larger than what we document. The contexts under which discrimination might take place in schools are likely to be subtle, because those who deliberately discriminate may not want to call attention to this behavior (taste discriminators), and differentials may emerge from the use of imperfect information (statistical discriminators) in ways that are more subconscious or based on implicit associations.

Importantly, we find that both White-Black score disparities and the presence of relatively low-performing Black students in classrooms where teachers start their careers strongly affect their evaluations of future student cohorts. Specifically, an increase of one standard deviation in the White-Black baseline performance gap among students in a teacher's initial classroom corresponds to teachers lowering their evaluation of subsequent Black students relative to White peers by an amount equivalent to over half of the estimated racial assessment gap. In essence, the worse Black students entering a teacher's initial classroom perform relative to White peers, the higher the assessment penalty for later cohorts of Black students compared to White classmates who do equally well on state-administered standardized exams. We conduct a series of falsification and robustness checks to ensure that teacher selection into classrooms based on hard-to-observe attributes do not explain these findings. Similarly defined performance differentials amongst future cohorts of stu-

[2]The most conservative estimates in Botelho et al. (2015) show that the magnitude of Black-White grading gaps is approximately 0.02 SD. Similarly, Hanna and Linden (2012) find that exams assigned to lower-caste children in India are graded at 0.03-0.08 SD below the exams of higher-caste children. Examinations of gender gaps show a magnitude of 0.05–0.25 SD (Lavy, 2008), while Lavy and Sand (2018) document a gender gap in math of 0.02 in middle school and 0.09 in high school. Even though gender gaps are distinct from racial differences, we believe this is also an important benchmark for juxtaposing effect sizes.

dents, for example, exert no effect on current teacher ratings of Black students relative to White peers. Our main estimates are also robust to the inclusion of teacher observable characteristics and initial school fixed effects interacted with the student race variable, suggesting that our conclusions about the effect of first impressions hold when only relying on within-school (cross-cohort) variation in early classroom composition. Taken together, our exercises indicate that teacher assessment behavior depends on the specific nature of first impressions. For example, the impact of early racial impressions only carries through to differential future assessments by race, and does not appear to influence teachers' evaluations of students by gender. Meanwhile, gender-specific performance gaps in initial classrooms only mattered for teachers' subsequent evaluations of girls vs. boys. We derive similar conclusions when examining the consequences of gender-specific racial gaps in early classrooms. Importantly, teachers' assessments of later cohorts of Black vs. White males only depend on the initial-classroom racial gaps among boys, not among girls (and vice versa).

The intensity of racial differences in teacher evaluation is particularly sensitive to the performance of Black and White students at the *bottom tail* of the first classroom's ability distribution. Teachers exposed to initial classrooms in which more White students out-perform the lowest-scoring Black student adjust their future evaluations of Black students downward. In contrast, teachers' relative assessments of later cohorts of Black students are not significantly responsive to early exposure to classrooms where high-performing Black students outscore a greater share of White students.[3] This asymmetry in teacher response prompts questions about why lower-scoring Black students are more salient or vivid than high-performing Black students in early classrooms. Notably, teachers respond when Black students' relative performance in initial classrooms adhere to negative stereotypes, but not when the high performance of those students defy prevailing stereotypes anchored on racial categories.[4]

[3]We interpret these results as being compatible with, yet not necessarily confirming, the presence of confirmatory bias in how teachers update their beliefs (Rabin and Schrag, 1999). This form of cognitive bias leads individuals to assign more weight to "preferred" beliefs, which in our context translates to novice teachers misreading the sequence of signals of Black vs. White students as supporting (society-wide) performance stereotypes. Evidence that contradicts prevailing stereotypes, such as high-performing Black students, is assigned less weight.

[4]In the education context, certain racial groups are associated with low academic achievement (Steele and Aronson, 1998; Alesina et al., 2018). The existence of a negative stereotype characterizing African Americans and low academic performance enable the mention of race to impair the performance of otherwise high-achieving Black students (Steele

Understanding teacher evaluation differentials is an important endeavor considering its potential contributions to the well-documented persistence of racial gaps in human capital (Neal, 2006; Reardon and Robinson, 2008; Clotfelter et al., 2009).[5] There is evidence that teacher expectations shape student achievement and the propensity to steer students towards gifted and talented education or particular fields of study (Donovan and Cross, 2002; Lavy, 2008; Burgess and Greaves, 2013; Botelho et al., 2015; Lindahl, 2016; Papageorge et al., 2016; Card and Giuliano, 2016; Lavy and Sand, 2018; Lavy and Megalokonomou, 2019). As a result, teachers who differentially assess their students by racial or ethnic group can exaggerate the sorting of students into various academic tracks, perpetuating existing gaps and exacerbating within-school segregation (Clotfelter et al., 2020).

There are also likely indirect consequences of teacher differential evaluations. Evidence shows these can become self-fulfilling prophecies by affecting parents' and students' own beliefs and behaviors (Rosenthal and Jacobson, 1968; Jussim and Harber, 2005; Hill and Jones, 2017), and can ultimately lead to changes in skill investment decisions. That is: if children's perceived competence increases the returns or reduces the costs of investments, as in the traditional human-capital framework (Becker, 1993), this feedback mechanism can reinforce racial gaps in the accumulation of human capital.[6] As a result, teacher evaluation differentials that make their way feedback to students and parents may lead to gaps in attainment, school choice, future scholastic performance and, ultimately, occupational choices and labor market outcomes (Mechtenberg, 2009; Lundberg and Startz, 1983). Efforts to bridge racial gaps in achievement and attainment can therefore benefit from a more informed understanding of this input and its relation with a teacher's early-career

and Aronson, 1998). Alesina et al. (2018) shows systemic teacher bias against immigrant students in grading. The authors trace the source of these racial differences to stereotypes using results from Implicit Association Tests (IAT).

[5]Longitudinal studies furthermore show that disadvantages among Black students emerge during early childhood and persist or grow throughout the schooling years. See Phillips et al. (1998), Hedges and Nowell (1999), and Reardon and Robinson (2008). Cautionary notes on these findings can be found in Bond and Lang (2013). Equivalent discussion on Hispanic-White gaps can be found in Reardon and Galindo (2009), for example.

[6]Dizon-Ross (2019) shows results of this mechanism by randomizing transcript information to parents. In her Malawi context, providing parents with performance information caused them to increase the school enrollment of their higher-performing children and to decrease the enrollment of lower-performing children. See also Papay et al. (2016). Fortin et al. (2015) find that gender differences in post-secondary expectations are the most important factor accounting gender gaps in school performance.

trajectory.

2 Related literature and contribution

Teachers are widely acknowledged to be a key input into education production and student learning (Chetty et al., 2014). Their interactions with students are increasingly scrutinized as a meaningful source of influence on student performance. One important way in which teachers can shape student outcomes is through grading or other assessments. Early studies in sociology identify teacher bias as a factor in course grading in the United States (Sexton, 1961; Rist, 1973; Farkas et al., 1990). Work following these early contributions has uncovered mixed evidence.[7] There is also a considerable number of contributions from the social psychology literature focusing on teacher's perceptions of Black and White children (see Ferguson (1998, 2003) and references therein), which again only unveils weak relationships between stereotypes and measures of discriminatory actions.[8]

Recent studies in economics, on the other hand, largely document significant race and gender differentials in teacher expectations and grading. For example, Figlio (2005) uncovers evidence of lower teacher expectations for those perceived to have African American ancestry, even after controlling for performance in standardized exams.[9] A common approach is to juxtapose subjective teacher evaluations with blind assessments of student performance. One set of papers capitalizes on the fact that students in Israeli high schools take two examinations covering the same material with the same format during senior year, and that the grading of each exam happens under different anonymity regimes. Using the blind score as the counterfactual to the non-blind teacher score, Lavy (2008) finds evidence of discrimination against males. Teacher biases based

[7] Large- (Williams, 1976; Sewell and Hauser, 1980) and small-scale scale empirical studies (Natriello and Dornbusch, 1983; Leiter and Brown, 1985) in that field do not detect significant biases on the basis of factors such as race, gender, and social class.

[8] See review of studies in Macrae et al. (1996). DeMeis and Turner (1978), unlike most of this literature, find significant discrimination against Black students in an experimental setting.

[9] Similar findings are present in audit-like studies. Hinnerich et al. (2011) transcribe and blindly re-grade tests assessed by teachers in Sweden and estimate gender (insignificant) and nationality (significant) gaps. A similar exercise conducted in Germany by Sprietsma (2013) also uncovers biases against exam solutions which had Turkish-sounding names randomly allocated to them (relative to German-sounding names).

on class-level gender differences furthermore have both short and long-term consequences for boys' and girls' human capital accumulation (Lavy and Sand, 2018; Lavy and Megalokonomou, 2019).[10] Blind/non-blind contrasts are also explored in a randomized control trial designed and implemented by Hanna and Linden (2012). The authors identify statistically significant positive differences between blinded and non-blinded scores for members of lower castes in India (relative to upper castes), which is clear evidence of discrimination. Finally, Burgess and Greaves (2013) and Botelho et al. (2015) use large-scale observational data in the UK and in Brazil, respectively, to investigate differences in teacher grading according to ethnic/racial background. They juxtapose objective tests with subjective teacher assessments and document significant underassessment of Black pupils (Black Caribbean and Black African in the case of the UK).

Our study builds on this literature by employing both blind and non-blind assessments of student mastery over the same skill set. In light of previous discussions, we underscore the contributions of our study context. First, we use large-scale observational data from the United States that provides plausibly objective measures of student math and reading mastery alongside subjective teacher assessments of the same underlying skillset evaluated on the same scale. Therefore, our blind and non-blind measures are well-suited for the task at hand, as both measures are taken contemporaneously and teachers are explicitly instructed to evaluate skill mastery over considerations of student behavior which may well be factored into the assignment of grades. We see this as an important advantage of our design relative to those that employ teacher assessments in the form of actual grades. While our juxtaposition of teacher assessments and standardized test scores aims to capture evaluation bias, we acknowledge that this measure stops short of fully exploring racially biased behaviors of teachers embedded in the very test scores that anchor our models. These may include teachers' varied treatment or mentoring of students across racial groups in a manner that differentially influences students' End-of-Grade test scores. In doing so we follow the literature in juxtaposing blindly and non-blindly graded scores to identify evaluation biases.

[10]Terrier (2020) similarly shows teacher favoritism towards girls using blind and non-blind test scores, and finds that girls as a consequence are more likely to choose a high school science track. Avitzour et al. (2020) probed the origins of these biases and document a correlation between *implicit* gender stereotypes and teacher assessment behavior.

Second and perhaps most importantly, we rely on detailed longitudinal information for both students and teachers to closely examine the central question of our work: the role of first impressions. The design of our empirical analyses follows the statistical discrimination and learning literature in conceptualizing teacher assessments as the weighted combination of imprecise ability measures and a prior based on racial group membership.[11] According to this framework, teachers who aim to evaluate a current Black student will combine noisy measures of the student's performance with the first impression they had about Black students' relative ability in their initial classroom. In essence, we hypothesize that the types and profiles of students teachers face in their initial classrooms may shape their expectations in subsequent years. This hypothesis is grounded in the recognition that in a sequence of signals, the first piece of information to which one is exposed is particularly salient and can shape beliefs about individuals and groups (Asch, 1946; Rabin and Schrag, 1999; Ambady and Skowronski, 2008).[12]

This conceptual framework lends itself well to investigations regarding learning, such as the seminal work presented by Altonji and Pierret (2001) in the labor markets and employer learning context. Increases in the signal-to-noise ratios arise mechanically from continuous interactions between the same employer and employee, since average productivity evaluated over time is less subject to measurement error than any single measure. The analogous education context for assessing learning involves repeated teacher interactions with the same student. The model predicts that as teachers get better at measuring the student's subject-specific knowledge, the role of race-based priors should diminish. We observe that our elementary education context is distinct from this form of learning from repeated interactions, as teachers do not typically observe the same students across multiple years.[13] Instead, our conceptualization corresponds to a process of *updating*, in which priors regarding specific demographic groups become posteriors once the teacher interacts with new cohorts of students belonging to the same demographic groups. Thus, teacher experience

[11] See, for example, the representation in Aigner and Cain (1977).

[12] We focus on first impressions given the extant theoretical literature and the salience or vividness of the first relative to subsequent impressions. It does not necessarily follow that impressions from second or third classrooms are irrelevant. We empirically evaluate the consequences of later exposure as an additional check.

[13] We acknowledge that teachers have repeated interactions with the same students within an academic year, but we lack the necessary intra-year data on student performance and assessments to investigate learning in this setting.

gained by instructing additional cohorts is a form of updating rather than learning about a particular student's ability over time in the manner of Altonji and Pierret (2001).

These differences notwithstanding, we retain key theoretical components from the statistical discrimination and learning literature to conceptualize teacher assessments, while contributing evidence that priors regarding the *current* cohort of Black students are correlated with the teachers' *first classroom impressions* of Black students' relative performance. Rich data on course membership and linked teacher-student information enable us to examine if first classroom attributes – such as the average performance of incoming Black or White students or the race/ethnicity of students at the extremes of the performance distribution – affect teacher assessments of future students belonging to the same racial group. Research exploring the origins of racial differentials are limited. Our study examines these issues in K-12 education with a focus on the extent to which racial evaluation differentials are influenced by teachers' early career experiences.

3 Data and descriptive statistics

3.1 North Carolina administrative data

We use administrative data on students, teachers, and course rosters from the North Carolina Education Research Data Center (NCERDC) to examine racial differentials in teacher assessments and the effects of initial classroom experiences. Individual and teacher identifiers enable the linking of teachers' demographic attributes, work experiences, subjective assessments of students' skills, initial classroom compositions, and students' characteristics, and blind-scored test performance.

In order to identify novice teachers, we use years of experience as indicated by teachers' pay grades. Legislated salary schedules in North Carolina set salaries according to education level and years of experience. We designate novice teachers as those with zero years of experience teaching for the first time in either a fourth or a fifth grade classroom. Novice teachers thus defined are cross checked with personnel files that denote when an individual enters their first year of educational employment. Teacher-level data also provide information on teacher gender, race, educational

background, and licensure history that we utilize in our analyses.

Next we use course membership data to characterize multiple dimensions of new teachers' initial classroom (IC) experiences. In particular, we use race composition and achievement information from the first cohort of students faced by novice teachers to construct measures of each teacher's IC conditions. These first impressions are based on test scores taken during the prior school year, *before* students interact with the teacher in question. For example, the first impression for a fourth-grade novice teacher who starts her career in the 2009-2010 academic year is based on her students' third grade End-of-Grade test scores from the 2008-2009 school year. We rely on lagged test scores because a given teacher cannot influence this measure of ability for her current set of students. Note that this precludes the inclusion of third grade teachers in our main analysis sample, since students are not taking standardized exams before that grade within the North Carolina system. We use these raw measures to compute group-specific summary statistics such as average standardized scores for each racial group. These variables, together with the shares of under-represented minority students by class, capture student composition and baseline ability distribution in each teacher's IC.[14] We match this information to the list of novice teachers and retain observations with non-missing IC characteristics. These IC measures are then linked to data on the test scores and teacher assessments of fourth and fifth grade students that these novice teachers face *after* their first year on the job.

To contrast teacher assessments of student abilities and students' actual performance, we rely on NCERDC data between 2007 and 2013 because these are the only years in which we have both End-of-Grade and teacher assessment data for both subjects. EOG tests aim to measure student proficiency at each grade level and are used in calculations of school performance under state and federally mandated programs. They consist of multiple-choice questions administered during the last three weeks of the school year. Each answer sheet is scanned and scored using software

[14]Classroom membership information is only included on NCERDC data starting in 2006. Therefore this imposes a binding restriction on the sample of teachers for which we can know classroom composition in their first incursion in the system. Of the unique teachers who started in 2006 or later and have between 1-3 years of experience, slightly more than half had non-missing initial classroom attributes. The majority of teachers who were missing early classroom variables worked in grades other than 4 or 5 during their first year. When we compare the demographic characteristics of teachers who had or were missing early classroom attributes, we find minor and mostly insignificant differences.

provided by the state Department of Public Instruction. The raw scores are also mapped to a 1-4 achievement level scale denoting insufficient mastery, inconsistent mastery, consistent mastery, and superior performance, respectively.[15] Since EOGs are machine-scored using a common rubric, we consider these assessments of math and reading ability as "blind" with respect to the racial identity of students.

During these same years, instructor questionnaires accompanied EOG tests designed to measure student proficiency. In these, teachers were required to provide their assessment of each student's achievement level (1 to 4) for math and reading comprehension at the same time students are undertaking the examination. According to the Department of Public Instruction, these assessments are used as an average across all teachers at the state level to calibrate the translation of the continuous Item Response Theory score distributions into the aforementioned four-point scale. We understand these subjective evaluations are not used by administrators as inputs in teacher performance assessment by principals and school-district administrators. Therefore, we also believe teachers have no incentive to be untruthful in their assessments. Importantly, the instructions ask them to identify each student who *"in the [subject] teacher's professional opinion, clearly and consistently exemplifies one of the achievement levels listed."* Moreover, teachers are explicitly told to focus *solely* on mastery over considerations of student behavior.

We restrict data to elementary school teachers because they usually interact with the same group of students across subjects rather than teach the same subject across multiple classrooms. This prolonged exposure ensures that they should be familiar with students' mastery of both math and reading. The fact that teachers know which student they are evaluating and the race and

[15] Throughout this study, the detailed description of each achievement level is as follows:

1. Students performing at this level do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level.
2. Students performing at this level demonstrate inconsistent mastery of knowledge and skills in this subject area and are minimally prepared to be successful at the next grade level.
3. Students performing at this level consistently demonstrate mastery of grade level subject matter and skills and are well prepared for the next grade level.
4. Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade level work

ethnicity of each student renders their assessments "non-blind." They provide assessments before knowing the students' actual test results. As such, the two measures of student skills are taken contemporaneously: machine-scored standardized tests and teachers' assessments of each student using the four-point achievement scale of insufficient, inconsistent, consistent, or superior mastery. Both measures are captured in the Spring of each academic year. Notably, our juxtaposition of teacher assessments and student test scores do not take place during the first year of a teacher's career. This particular cohort of students is only used to construct measures of a first impression. To return to the above example, a novice teacher who begins in the 2009-2010 academic year forms a first impression based on her then students' lagged test scores at the end of the 2008-2009 year, while we only examine evaluation biases for new student cohorts in 2010-2011 and beyond using contemporaneous teacher assessments and test scores from those subsequent years.[16]

The final dataset includes elementary students in a mixed-race fourth or fifth grade class who were taught by teachers with 1 to 3 years of experience (we classify novice teachers as having 0 years of experience). In analyzing the role of first impressions, we further limit observations to students taught by teachers with non-missing data on IC conditions. In all, our most comprehensive analytic sample includes 2,196 teachers across 7,494 unique classroom-subject groups, amounting to 203,062 student-classroom-subject level observations. A subset of these observations, for which we recovered information on a teacher's initial classroom experience, is used for the analysis of the relevance of first impressions for teacher evaluations of current students.

3.2 Descriptive statistics

Table 1 details student, classroom, and teacher characteristics. We restrict our sample in this table and subsequent analyses to classrooms that serve both Black and White students. In columns (1a) and (1b) we present descriptive statistics for the full sample, while columns (2a) and (2b) reproduce the same indicators for the sub-sample for which we have detailed information on teachers' initial classrooms. We observe minimal differences between the full and sub-samples across all data

[16]Diagram OA1 in the Online Appendix describes this data structure in detail.

dimensions. Students in both samples (Panel A) are just under 11 years of age, representing a nearly even split between 4th and 5th graders as well as between math and reading classes. The samples are also balanced between boys and girls. In terms of racial composition, 47% of the observations are non-Hispanic White and 31% are non-Hispanic Black. The average student misses 6 days of classes during the school year, with 4% who are chronically absent due to missing more than 10% of the school year. On average, students score about a third of one standard deviation above the official proficiency cutoff in the state's EOG test, corresponding to 67% being proficient according to the same standards. We also see that teachers evaluate the average student at higher levels than the one obtained in standardized tests (2.8 versus 2.7 along the four-point scale), which correspond to a 2 to 3 percentage point difference in proficiency rates.

Since these differences between teacher and EOG-based evaluations form our core focus, we present a more detailed juxtaposition using our sample of classes led by teachers with one to three years of experience in Figure 1. We compare subject-grade-specific teacher evaluations on the four-point scale on the vertical axis with corresponding End-of-Grade standardized scores on the horizontal axis. That is to say, we nonparametrically analyze the bivariate relationship between contemporaneously measured End-of-Grade test scores and teacher assessments, separately for Black and White students. We also portray the estimated density of EOG scores underlying these two-way relationships. The density functions show a substantial score advantage among White students relative to Black students across the performance distribution. These depictions also indicate that teacher assessments tend to favor White students across all levels of performance, despite the strong relationship between their evaluations and standardized test scores. We formalize and quantify these findings using our econometric specification below.

Panel B of Table 1 presents classroom-level information across our samples. We observe that nearly 9 out of every 10 elementary school teachers are female. The racial makeup is predominantly White, with 91% of teachers in this category while 7% are Black. The clear skew in the sample towards White and female instructors is consistent with demographics of the national teach-

ing labor force.[17] The racial composition for an average classroom of 27 students is 45% White and 32% Black. In terms of matching between the racial composition of the student body and a teacher's race, we report in Figure OA1 that Black teachers tend to have classrooms with a larger share of Black students. There is, nonetheless, significant overlap between the distribution of the share of Black students in classrooms led by White and Black teachers in our sample. By construction, our classroom sample includes teachers with one to three years of experience (average of 1.78 years). Over 83% of teachers have a BA degree while 14% have completed post-graduate studies.

We then move to the central objective of the paper: to scrutinize the role of initial classroom experiences in shaping racial biases in subsequent teacher assessments. We begin with exposure, or whether a teacher had at least one Black student during their first year. Teachers who taught at least one Black student may rely on performance signals to update their priors about this group in ways that are meaningfully different from those who had no previous classroom contact with Black students. Among teachers who had both Black and White students in their initial classrooms, we examine the ability distributions of these two groups using a measure that precedes the interaction with the teacher in question (as described above). The extent to which the performance distributions of Black and White students diverge from one another can shape subsequent expectations and, therefore, assessment differentials of future cohorts of students. This is what we set out to measure.

We generate several measures to capture teachers' initial classroom experiences. Panel C of Table 1 presents statistics at the level of the novice teacher. In addition to verifying that demographics and educational background characteristics correspond to classroom-level observations, we provide additional descriptive statistics for the subsample with initial classroom information.[18] These show that teachers in our sample initiated their careers in classrooms with racial compositions similar to the ones they are currently assigned to. The next variable in Panel C is the

[17] According to the National Center for Education Statistics, 89% of public school elementary teachers in 2017-2018 were female, while 79% of both elementary and secondary school teachers were White.

[18] We require information on racial composition and lagged performance in standardized tests by both White and Black students.

White-Black test score gap in early classrooms. The measure averages over individual z-scores separately for White and Black students and takes the difference. The average initial classroom assigned to the teachers in our sample has Black students lagging behind White peers by 0.55σ.

While the White-Black test score gap in early classrooms may shape the formation of teachers' future expectations, we anticipate that other attributes of the performance distribution could also matter. In particular, elements that are vivid, concrete, and proximate may play a bigger role in shaping inferences and behavior (Nisbett and Ross, 1980). We operationalize vividness by focusing on outlier students, that is, students who are at the tails of the ability distribution who may stand out in teachers' memories. We capture the non-overlapping tails of the ability distribution by constructing measures for the share of White students whose incoming scores are below (above) the lowest-scoring Black (highest-scoring) student to describe the extent to which one tail underperforms the other.[19] We compute that 33% of White students in the average teacher's initial classroom scored above the highest-performing Black student based on their previous EOG test performance (or lagged score) while 9% of White students scored below the lowest-performing Black student.

Descriptive statistics on early classroom conditions faced by the teachers in our sample disguise substantial heterogeneity by White and Black students' incoming levels of preparation. Panel A of Figure 2 shows the gap in scores between White and Black students in initial classrooms. While the relative rightward placement of the distribution demonstrates that the average gap favors White students, we observe substantial variation across teachers. Panel B breaks out White and Black students' relative performance at the initial classroom level, with the 45-degree line indicating classrooms exhibiting score parity across White and Black students. The presence of many classrooms along this line indicates a wide range of student performance for any given White-Black score gap (zero in this case). Observations below (above) the line involve initial classrooms where White students outperform (underperform) Black students. The final panel examines initial classroom contexts faced by the teachers in our sample at the intersection of race and gender, by juxtaposing

[19]These measures are akin to the representativeness heuristic discussed in Bordalo et al. (2016) and Gennaioli and Shleifer (2010) in the context of stereotype formation.

racial gaps among students of the same gender. The figure demonstrates that even within the same classroom, the White-Black achievement gap vary substantially from one gender group to another (i.e., Black female students outperform White female peers while Black males under-perform their White male counterparts, or vice versa). Below, we capitalize on this variation in initial classroom conditions to investigate the relationship between gender-specific racial differentials in early classrooms and teachers' evaluations of current students.

4 Empirical approach

4.1 Research design

We first describe our approach towards examining the influence of race on teacher assessments, before turning to the role of first impressions on teachers' subsequent student evaluations. To determine the extent of racial differences, we juxtapose teacher assessments with students' contemporaneous performance on standardized End-of-Grade exams. We advance that both assessments measure the same underlying skills, and any remaining racial gaps in teacher evaluations not accounted for by test performance are also not explained away by differences in student subject-specific cognitive skills and demographics.

Two contextual pieces of evidence support our contention that teacher assessments capture student mastery in math and reading, rather than other cognitive or socio-emotional skills. First, teachers are given explicit instructions to focus on mastery in the tested subject over performance in other subjects or student behavior. The prompt states that *"The [subject] teacher should base this response for each student solely on mastery of [subject]. The [subject] teacher may elect to use grades as a starting point in making these assignments. However, grades are often influenced by factors other than pure achievement, such as failure to turn in homework. The [subject] teacher's challenge is to provide information that reflects only the achievement of each student in the subject matter tested."* Our interpretation of these instructions is that teachers are asked to offer a second-

opinion on the same math and reading skills evaluated by EOG tests.[20] The emphasis on "the subject matter tested" suggests that teachers did not read these instructions as intending to capture performance in dimensions beyond what the EOG test measures.

The second consideration is the close link between EOG tests and inputs into teacher assessments in an accountability-based system that emphasizes test preparation. We interpret teachers' ratings on the four-point scale as deriving from a sequence of performance signals that the teacher receives about a given student over the academic year. Many of these inputs involve in-class assessments modelled after EOG exams, sometimes distributed by the state Department of Public Instruction prior to the testing period. As such, the contents of our measure of teacher assessments are directly informed by EOG test preparation resources, suggesting that they should measure the same underlying skills. Under this assumption, variables such as race should not systematically influence teacher evaluations after holding constant performance on EOG exams. Hence, we view differences across racial and ethnic groups as indicative of differential assessment and the influence of priors.

Notably, the four-point achievement scale used in teacher assessments aligns with state-approved curriculum standards. This suggests that teachers should be rating students based on externally-determined expectations rather than local reference points such as the classroom. We emphasize that these notions of absolute performance are clearly communicated to teachers by education authorities.[21] This reliance on an absolute performance scale alleviates concerns about reference bias, in which teachers form subjective assessments based on relative cross-student comparisons (Elder and Zhou, 2021). Another element of our setting that is advantageous involves our use of classroom-subject fixed effects, in contrast to within-school variation used in Elder and Zhou (2021). This approach enables us to absorb classroom-specific biases, such as optimistic teachers uniformly inflating students' ratings, so as to only rely on the localized context of within-class variation.

[20]Teachers provide these subjective assessments during the testing period for EOGs, and their aggregate responses are used at the state level to calibrate achievement levels for the exam.

[21]The North Carolina Department of Public Instruction collects teacher evaluation data as part of regular testing procedures.

Even as these considerations address reference bias, we present additional evidence that our context is less prone to these concerns. First, we follow the reasoning presented by Elder and Zhou (2021) and provide additional descriptions of the variation contained in our data.[22] Table OA1 indicates that differences in the contribution of within-classroom variation in subjective and objective measures of performance are closer to each other in our sample than in the sample reported in Elder and Zhou (2021)'s Table 4. This reassures us that we are unlikely to be affected by this form of reference bias. Second, Figure OA2 shows that classroom racial composition and subjective and objective performance measures follow each other more closely in our study context than in Elder and Zhou (2021)'s Figure 1, Panels A, B, D and E.[23] Most importantly, we show that the Black-White gap in both subjective and objective evaluations in our sample exhibit nearly a parallel pattern across different classroom compositions. These are indications that our sample does not suffer substantially from the potential reference biases raised by the authors in the context of ECLS-K.

4.2 Empirical specifications

We begin by examining the relationship between race and deviations in teacher assessments from contemporaneous EOG performance. This corresponds to the following formulation:

$$c_{irst} = \alpha_1 Black_i + \alpha_2 f(scores_{irst}) + X_{irst}{}'\boldsymbol{\beta} + \eta_{rst} + \epsilon_{irst} \tag{1}$$

where c_{irst} captures teacher evaluations for student i under teacher r for subject s and time t. Our main outcome of interest uses teacher assessments on the four-point scale.[24] Crucially, we condition on a flexible polynomial function of End-of-Grade test performance, $f(scores_{irst})$. X_{irst} includes indicator variables for other racial and ethnic groups in order to guarantee the interpreta-

[22] Additional discussion is presented in the Online Appendix, Section B.

[23] While both share a negative slope in our sample, the authors of that study report a positive slope between teacher evaluations and school-level shares of Black students, while showing a negative slope between objective test scores and the share of Black students.

[24] We also present alternative dependent variables based broadly on cardinal and ordinal scales: a) an indicator variable for attaining proficiency, expressed as an achievement level of at least 3 which is an absolute standard common across all public schools in the state, and b) an indicator for being rated above the class mean.

tion of the α_1 coefficient as "relative to White students". We also include demographic attributes such as gender and age in the vector of covariates. The model further includes teacher-subject-year fixed effects η_{rst} to account for differences in assessment practices and other hard-to-observe attributes that vary across classrooms at the teacher, subject, and year levels. For instance, this addresses aggregate racial differences induced by White students possibly disproportionately sorting into classrooms with more lenient teachers who inflate assessments uniformly across all students. Since elementary school teachers are assigned one class per year, identification is based on within classroom-subject variation. This empirical model underscores race as a key element affecting teachers' expectations of student mastery. We scrutinize the direction and magnitude of α_1 in relation to the dependent variable. A negative coefficient on Black_i indicates that teachers rate Black students lower relative to White classmates with the same test scores.

We subject our main specification to a number of robustness analyses, beginning with checks for common support across racial groups in test scores and sensitivity to different functional forms of $f(scores_{irst})$. We ensure that our estimates of Black-White differences in teacher assessments are not sensitive to the presence of non-overlapping portions of test scores' support. We also replace the parametric specification with a more flexible control of EOG fixed effects. Another set of analyses accounts for the possibility that teacher assessments are inclusive of other cognitive or behavioral attributes, despite explicit instructions otherwise. We estimate models augmented with student absences and suspensions as proxies for student misbehavior, engagement and self-reported student effort in school-related activities, and the previous teacher's evaluation on the same four-point scale. To the extent teacher assessment embeds information on student cognitive ability beyond the included achievement and behavioral characteristics, this lagged assessment measure helps account for such characteristics not observed by the econometrician.

The second and central part of our empirical strategy relates the extent of racial differences in teacher assessments to early classroom experiences among novice teachers. We extend the model to capture differences in teachers' initial classrooms by allowing the parameter α_1 to be a function of those early experiences (IC_r). Initial classroom measures include whether (and the intensity

by which) novice teachers were exposed to Black students in their first classrooms and the nature of such exposure as summarized by the magnitude and sign of average White-Black test score differences, and non-overlapping tails of White and Black score distributions within those first classrooms. We estimate the following specification:

$$c_{irst} = (\alpha_{10} + IC_r{}'\alpha_{11})Black_i + \alpha_2 f(scores_{irst}) + X_{irst}{}'\boldsymbol{\beta} + \eta_{rst} + \epsilon_{irst} \tag{2}$$

Teacher assessments c_{irst} and student EOG scores $f(scores_{irst})$ are measured 1 to 3 years *after* teachers' first year, while IC_r describes early classroom conditions in year 0. α_1 coefficients collectively quantify the size of racial differences in teacher assessments within the contemporaneous classroom after adjusting for EOG test scores, student demographics, and classroom-subject fixed effects. Our coefficient of interest is α_{11}, which captures heterogeneity in racial differences by teachers' initial classroom experiences.

To alleviate concerns that novice teachers may select into classrooms based on unobserved characteristics, we undertake a number of falsification checks and additional analyses. First, we examine whether teacher assessments are influenced by exposure to *future* classroom conditions defined in the same manner. Significant results here can indicate the systematic placement of teachers into classrooms that are confounding identification, while null findings provide assurance that our results are not due to selection generating a relationship between contemporaneous teacher ratings and initial or future classroom attributes. We then estimate fully stratified samples based on the sign of the initial classroom White-Black test score gap, and explore whether racial differences in early classroom performance affect teachers' subsequent *gender* assessment gaps. Analogously, we examine whether initial classroom conditions defined by gender differences affect teachers' later assessments of Black vs. White students. An additional and informative set of analyses rely on substantial variation in the magnitude and direction of gender-specific racial gaps in initial classrooms to examine whether first impressions of early racial score gap for boys (girls) persevere to differentially affect teachers' assessments of boys (girls) later on. Finally, we extend the model to account for a set of teacher attributes denoted by T_{rt}:

$$c_{irst} = (\alpha_{10} + IC_r{}'\alpha_{11} + T'_{rt}\alpha_{12})Black_i + \alpha_2 f(scores_{irst}) + X_{irst}{}'\boldsymbol{\beta} + \eta_{rst} + \epsilon_{irst} \qquad (3)$$

The observable teacher attributes that we account for using T_{rt} include gender, race, educational attainment, licensing and years of experience. We also augment the model to include a full set of teacher's initial school fixed effects interacted with $Black_i$. Their inclusion implies that the estimation of α_{11} is based on *within-school* variation in first classroom assignments. The underlying assumption is that among the pool of novice teachers, their early classrooms in a given school are not systematically assigned based on a predisposition for racial bias in assessment. We assume that administrators have no direct way of inferring race-related attitudes among novice teachers hired by the school.[25]

We show corroborating evidence for this assumption by examining the relationship between novice teachers' observed characteristics and initial classroom characteristics with and without school fixed effects. Online Appendix Table OA2 presents evidence that this strategy can aid the identification of first impression effects. Panel A employs variation within and across schools, while Panel B reports the ones corresponding to within-school variation only. Panel A shows that teacher demographic characteristics are sometimes associated with initial classroom racial composition and performance by racial group. These relationships become insignificant when conditioning on initial school fixed effects (Panel B).[26]

While this design is internally valid, we caution that our findings cannot necessarily be extrapolated beyond the population of novice teachers. Due to data limitations presented above, we have no direct way of assessing the impact of early experiences over more experienced teachers'

[25]Novice teachers may be more likely to be allocated to hard-to-staff schools. Indeed, evidence described in Clotfelter et al. (2006) indicates that highly qualified teachers tend to be matched with more advantaged students. While this is an important consideration, since we focus our analysis on a pool of novice teachers only, this pattern may affect the external validity of our findings but should not pose a threat to the internal validity of our estimates.

[26]It is important to consider literature on parents and administrators selecting teachers within a school (Kalogrides and Loeb, 2013). However, these previous studies focus on the contrast between novice and more experienced teachers for which reputation is likely the main driver of parental/principal's decisions. We believe that once we restrict to only novice teachers in the first year of their careers, there is simply no information on track records for these actors to use and, consequently, induce sorting.

evaluation patterns, or to test for learning in a context where teachers may statistically discriminate based on race (Altonji and Pierret, 2001). We can estimate, however, if those first impressions are still relevant up to three years afterwards.

5 Evidence on race-based differential assessments

5.1 Extent of assessment differentials

Table 2 begins formalizing the comparison of means we illustrated with Figure 1 above. On the four-point scale, teachers rate White students at 3.055, relative to 2.598 for Black students. The raw difference of 0.457 reflects a myriad of factors, including differences in ratings behavior across teachers and actual differences in student mastery. We add classroom-subject fixed effects to account for the former, yielding an attenuated gap of 0.416, which in principle reveals that the bulk of the differences in evaluation between Black and White students does not come from differential exposure to more rigorous or lenient teachers (at least not in this sample of mixed-race classrooms). The inclusion of normalized test scores (entered as a fourth order polynomial) leads to an adjusted Black-White gap of 0.059. The magnitude of this Black-White gap remains stable and becomes 0.060 after adding demographic controls on gender and age relative to same-grade peers to a model that already includes class-subject fixed effects and normalized test scores. The estimation also reveals that the relationship between EOG test scores and a teacher's evaluation of mastery is strong, as predicted by our conceptual reasoning above. The marginal effect of a one standard deviation increase in blind-scored EOG performance, evaluated at the proficiency cutoff, is equivalent to increasing teacher assessments by 0.68 on the four-point scale. The estimate also aids our interpretation of the racial differential – the measured racial gap in evaluation of 0.060 is equivalent to what a 0.09 standard deviation reduction in EOG performance would produce.

Table 2 also presents two alternative representations of the teacher rating gap. The first transforms the four-point scale into an indicator variable for reaching proficiency, or an achievement level of at least 3. Column 6 shows that, conditional on EOG test scores and other covariates em-

bedded in the previous specification, Black students are 2.5 percentage points less likely to be rated as proficient than their White peers in the same classroom, which is equivalent to a 0.05 standard deviation difference. As another point of comparison, teacher rate 79% of White students and 58% of Black students as proficient in the full sample. Column 7 shifts the representation of teacher evaluations from an absolute standard based on achievement levels to an ordinal scale involving relative classroom comparisons. The dependent variable is an indicator for being rated above the class mean. Black students are 3.1 percentage points less likely to be rated as such relative to White classmates with equal performance (the baseline average among Blacks is 43%).

Given the importance of EOG scores in our conceptualization and empirical specifications, we undertake a series of analyses to ensure its robustness to functional form assumptions and to ensure common support between Black and White students in the same class. Online Appendix Table OA3 shows that there is substantial overlap between the EOG score distributions of Black and White students. If we trim the non-overlapping tails of Black and White students' score distributions, the resulting racial gap is almost unchanged (Column 1). The next specification takes this even further by only retaining observations for which a Black student has a White classmate with the exact same EOG score. The coefficients are almost identical. Finally, we use EOG fixed effects in place of a fourth order polynomial, and the stability of results shows that our estimated racial disparity is robust to the alternative semi-parametric specification.

We then examine the robustness of our main model to additional socio-demographic and behavioral attributes that may be correlated with both race and teachers' notions of academic competency. Table OA4 begins with the same specification shown in Column 5 of Table 2 and a Black-White evaluation gap of 0.060. The second column shows a Black-White gap of 0.053 after adjusting for days absent and and lagged teacher evaluations on the same four-point scale. We include student absenteeism to account for the possibility that teachers may rely on behavioral attributes as inputs into their judgment, despite explicitly instructed by the Department of Public Instruction to do otherwise. To the extent that the lagged evaluation variable embeds some racial differences attributable to the previous teacher's bias, we may be underestimating the scope of

racial differentials with its inclusion in the model. The relative consistency of our results suggest these factors do not play a major role. The next pair of columns restrict the sample to years with suspension data and finds that the inclusion of days absent and suspended leaves a Black-White evaluation gap of 0.065. The last two columns restrict the sample to 2008-2012 to accommodate the limited availability of two time use variables summarizing the hours spent on homework and free reading. These proxies for effort and motivation do not dramatically affect the Black-White teacher evaluation gap nor alter our main conclusions.

The analyses so far pool observations across institutional, teacher, and student characteristics. Online Appendix Table OA5 stratifies the sample by subject, grade, and teacher experience. Teachers rate Black students 0.079 points lower in reading compared to 0.039 points lower in math. The relative precision of these estimates points to a larger differential in reading. Results echo findings in related studies documenting bigger effects in English relative to math (Lavy, 2008; Burgess and Greaves, 2013). In contrast, we find no evidence that racial disparities in teacher assessments are driven by particular grades or years of teaching experience. Table OA6 segments the pooled sample by student gender and age relative to the within-grade mode. Standard errors are sufficiently large that we are not able to reject the equivalence of coefficients corresponding to the subgroups by either gender or age.

5.2 First impressions

After establishing Black-White evaluation gaps using the full dataset, Table 3 evaluates the influence of initial classroom conditions on teachers' subsequent rating behavior. The first column replicates the adjusted Black-White rating gap of 0.060 points in Column 5 of Table 2 using the pooled sample. The next specification restricts to teachers whose initial classrooms contained at least one Black and one White student and had non-missing initial classroom information. Beginning with Column 3, we focus on how the extent and content of exposure to Black students during teachers' initial year shapes subsequent assessments.

The first specification examines whether teachers are more likely to attenuate racial disparities

in future student ratings when Black students make up a greater share of their first classrooms. The coefficient on interaction term is not significant, suggesting that teachers with a greater share of Black students relative to the sample mean in initial classrooms do not exhibit measurable differences in future White-Black assessment gaps. In Column 4 we turn to the nature of that initial exposure to Black and White students. The specification includes an additional interaction term between an indicator for Black and the White-Black EOG test score gap in the first classroom (centered at the sample mean).[27] For every standard deviation increase in the initial classroom White-Black test score gap, teachers tend to assess their current White students by an additional 0.033 points more than their Black students, even when these students have the same EOG test scores. This effect is equivalent to 55% of the observed Black-White assessment gap. The following specification focuses exclusively on the sign of the initial classroom White-Black test score gap. Column 5 shows that compared to teachers who did not have White students that outperformed Black students in their initial classroom, those that did end up having assessment gaps more unfavorable to current Black students by about 0.056 points. Over 15% of the sample had a teacher whose first classroom had higher-performing Black students, while the remainder had teachers who were exposed early on to higher-performing White students than Black peers. A closer examination that stratifies the sample by student gender shows that this pattern of subsequent teacher behavior pervades the assessment of both girls and boys, such that "penalities" do not accrue disproportionately to one gender group (Table OA7).

We interpret these results as the consequence of initial classroom exposure, and subject the findings to a series of placebo tests and robustness checks. To begin with, we examine whether exposure to *future* classroom conditions generates qualitatively similar findings. An answer in the affirmative would severely undermine our conjecture of causal effects and favor the interpretation of a selective allocation of novice teachers to classrooms. Table 4 limits the sample to observations with non-missing future classroom attributes calculated during the fourth year after teachers finish

[27] As detailed above, we use lagged scores to compute the Black-White average test score gap. For example, measures of initial classroom conditions for a new fourth grade teacher rely on End-of-Grade standardized tests in grade 3.

their novice academic year. We examine both the effects of initial (year 0) and future (year 4) classroom conditions on teachers' racial assessment gaps during years 1-3. While the extent of exposure to Black students early on does not influence teachers' subsequent ratings behavior (Column 1), teachers penalize future cohorts of Black students based on the magnitude of the White-Black score advantage in their initial classrooms. The magnitudes of their behavioral changes are statistically equivalent to those in Table 3, despite a significantly smaller sample. Strikingly, parallel findings on the effect of *future* classroom conditions show no accompanying differences in teachers' assessment behavior. This is expected if we are identifying the causal effects of classroom attributes, given that later classrooms should have no bearing on teacher ratings during years 1-3. The absence of significant effects also suggests no systemic relationship between teachers' racial differentials in student ratings and subsequent student compositions in future classrooms.

We further probe the lasting influence of early classroom conditions by estimating fully interacted models (stratified samples) based on the sign of the initial classroom White-Black test score gap. The estimation results in Panel A of Table 5 examine whether the nature of early exposure shape teachers' future evaluations of students not only by race, but also individual attributes such as gender. A notable finding from Table 5 is that relative White-Black performance in initial classrooms only strongly affects the differential assessments of current students by race, but has no influence on assessments based on gender 1 to 3 years later. Teachers for whom White students outperformed Black peers in the initial classroom reduce their evaluations of later cohorts of Black students by 0.048 points, compared to teachers who were exposed to early classrooms where Black students outperformed White students. In contrast, the sign of the initial classroom White-Black test score gap has no bearing on teachers' propensity to rate boys lower than their similarly-performing girl peers. According to our reasoning, racial gaps in initial classroom performance only provide relevant information for the formation of priors utilized for differentially evaluating current students according to racial identity.

Panel B in Table 5 replicates these analyses using the subset of observations with future classroom conditions. Column 3 shows differences in Black-White teacher assessments based on the

sign of the initial classroom White-Black test score gap that are statistically indistinguishable from the results above. We then replicate the analyses in Panel C by stratifying future classrooms into those with higher performing White or Black students, on average. Strikingly, the placebo test shows no differential effect on teacher ratings during years 1-3 based on this classification of future classroom conditions. This holds for both Black-White and female-male gaps in teacher assessments.

The evidence thus far point to teachers' assessment behavior being driven by the particular circumstances of initial classrooms. Panel D poses the question of whether these findings are specific to early *racial* gaps in performance and not generalizable to other demographic characteristics such as gender. We replace race-based initial classroom conditions with the female-male test score gap in teachers' first classrooms. Findings suggest that initial classroom gender gaps are relevant only for teachers' subsequent assessments of students by gender group. Teachers exposed to early classrooms with higher performing girls lowered their subsequent assessments of boys by 0.015 points, relative to teachers whose early classrooms had higher performing boys on average. Notably, gender-specific gaps in initial classrooms did not induce teachers to differentially assess Black vs. White students later on, suggesting that the specific content of those first impressions matter.

Further evidence in support of this claim is found in Table 6. We rely on substantial variation in gender-specific racial gaps evident in initial classrooms (Figure 2, Panel C) to examine whether teachers' evaluations of subsequent cohorts of Black males relative to White males are more sensitive to boys' White-Black relative score advantage in early classrooms, rather than the racial score gap among girls. Column 2 shows that teachers are indeed lowering future relative assessments of Black boys based on the early racial gap among boys only. An analogous specification for girls finds that teachers similarly reduce their subsequent relative ratings of Black girls when White girls in their first classrooms more strongly outperformed Black girls, but not when the White-Black score gap among boys was wider (Column 4).

Finally, to provide further assurance that we are documenting the consequences of initial class-

room impressions of racial differentials on future teacher assessments, we examine the robustness of our findings to accounting for teacher attributes that may relate to initial classroom conditions. Table OA8 begins by reproducing the coefficients in Columns 4 and 5 of Table 3 that assessments of future Black students decrease by 0.033 and 0.056 points when the initial classroom White-Black test score gap increases by one standard deviation (Panel A) or is greater than 0 (Panel B), respectively. Column 2 includes interactions between Black and teacher demographics (gender and race), background (has at least a Master's degree and was licensed in North Carolina), and years of experience. Notably, the inclusion of teacher characteristics barely changes the point estimates on the effect of the initial White-Black test score gap as measured by both magnitude and sign. This suggests that our estimate in Column 1 is not plagued by omitted variable bias from not accounting for sorting based on these teacher observables and their correlates. We report all coefficients on the teacher interaction terms to assess whether racial gaps in evaluation vary significantly across teacher attributes. Relative to teachers who are predominantly White, Black teachers have smaller racial gaps in future assessments, but this difference is not statistically significant. The sign of the interacted coefficient is consistent with previous literature on the advantages for Black students of having a racially congruent teacher who may hold higher expectations of student achievement and attainment (Dee, 2004, 2005, 2007; Gershenson et al., 2016). We furthermore do not observe significant differences by teacher gender or experience. All else equal, we find that teachers with at least a Master's degree and those licensed in-state have larger Black-White assessment gaps.

The final specification in Table OA8 includes a full set of initial school fixed effects interacted with the student race variable. As specified in the empirical strategy section, the model identifies the effect of first impressions using within-school variation in early classroom conditions. This specification reasons that novice teachers have little discretion in selecting particular classrooms within a given school. Column 3 shows that a one standard deviation increase in the relative White student score advantage in initial classrooms decreases teachers' assessments of later cohorts of Black students by 0.020 points. By the same token, we find that teachers who experience an initial classroom in which the average White student outperforms the average Black student tends to

evaluate mastery of current Black students at 0.039 scale points lower than those that experienced the reverse condition. The point estimates in Column 3 are significant and we cannot reject that it is the same as coefficients from models estimated without the interaction with initial school fixed effects.

The results so far underscore the consequences of first impressions as measured using average racial differences in test scores. Next we explore the different ways in which heterogeneity in average performance could materialize. For one, changes in average gaps could be translated into the relative movements of the distributions of Black and White students. Suppose we fix the White distribution and move the extremes of the Black distribution. Less overlap in these distributions implies greater statistical power in the test for Black-White differences in the initial classroom. Put differently, less overlap yields greater confidence among teachers of observing a difference. We operationalize this in Table 7 by examining students whose scores render them in the non-overlapping portions of the performance distribution.

Columns 1 and 2 replace the interacted term of White-Black score gaps with the share of White students who score above the highest- or lowest-achieving Black student in teachers' initial classrooms, respectively. The following specification includes both of these interactions. Corresponding results in Column 3 show that contemporaneous teacher assessments are significantly sensitive only to initial classrooms where the lowest-performing (but not highest-achieving) Black student under-performed relative to her White peers. When the lowest-scoring Black student underperform a larger share of her White peers in an initial classroom, teachers impose a larger penalty on current Black students. Specifically, the racial disparity in assessment for a teacher exposed to a first classroom in which the lowest-scoring Black student outscored 80% of the White students would be smaller than for a teacher exposed to a first classroom in which the lowest-scoring Black student outscored 20% of the White students. The difference between these two experiences would be equivalent to the average evaluation gap size of 0.060 points we estimate above. In contrast, there is no significant change in teachers' subsequent evaluations of Black vs. White students when we focus on variation in initial classrooms classified according to the performance of the

highest-scoring Black student relative to White classmates. Note that the coefficient still has the expected negative sign, indicating widening racial assessment gaps as a result of a larger share of White students outperforming the best Black student. These findings imply that teachers seem more reactive to comparisons between the worst students than between superstars.[28] We subject these results to another robustness check by including interactions of student race with teachers' initial school fixed effects. Corresponding point estimates in Columns 4-6 are statistically equivalent, providing further assurance that our results are not confounded by potential sorting of novice teachers across schools in their first teaching assignment.

Even though the asymmetrical influence of Black low performers relative to high-achieving Black students is consistent with a large body of research in psychology and a more recent strand of economics literature formalizing the intuition of confirmatory bias (Lord et al., 1979; Nickerson, 1998; Rabin and Schrag, 1999),[29] we cannot fully rule out updating based on a Bayesian framework in which teachers hold stereotypes against Black students early on but update their beliefs about that racial group as they interact with more students. The main reasons for this is that we only examine a relatively short period of teachers' careers of between one and three years after their first academic year. Teachers may not be able to gather sufficient data points about various racial groups during this compressed period.

A related question is whether impressions from the second or subsequent years also have enduring impact. Our results thus far indicate that first impressions matter (in the sense that initial classrooms influence the assessments of future student cohorts) but not that they are the only ones to matter. Efforts to answer this question face empirical challenges, however. For one, the compositions of initial classrooms for novice teachers are plausibly exogenous, particularly after condition-

[28] We also consider an alternative specification in which we define indicator variables for whether the highest (lowest)-scoring White student exceeds the highest (lowest)-scoring Black student within the teacher's initial classroom. Findings in Table OA9 confirm the overall interpretation that teachers are more responsive to racial differences among the lowest-performing students than those among high-achievers.

[29] Pioneering psychological studies demonstrate that in lab experiments, participants placed more emphasis on research that supported their own opinions and questioned research that countered their beliefs (Wason, 1960; Lord et al., 1979). They assign more weight to preferred beliefs, which inhibits their ability to arrive back at the correct hypothesis after a sequence of signals. In their seminal work, Rabin and Schrag (1999) repeatedly bring up a classroom example to illustrate this phenomenon, in which "teachers misread performance of pupils as supporting their initial impressions of those pupils."

ing on initial school fixed effects, while the sorting of more experienced teachers into classrooms may be subject to greater discretion. While we construct measures of the first impression using lagged scores of students before they ever encounter the novice teacher, the pre-interaction scores of the second cohort of students may have non-exogenous components. With these caveats, we undertake a corresponding analysis on the role of second impressions based on teachers' second year classrooms. Table OA10 shows significant impacts of first impressions but non-significant impacts of second impressions, even though standard errors do not allow us to affirm that the point estimates are indeed different from each other. These results do not suggest that second or third impressions are irrelevant, but they do underscore the salience or vividness of first impressions.

Examining the consequences of exposure to different cohorts also prompts questions on whether, and if so, how teachers *learn* about a given student's ability (not her racial group's) over time. The classroom context we study differs crucially in one respect from statistical discrimination and learning studies in the workplace: interactions between an elementary school teacher-student pair usually last for only one year, especially in elementary schools. As such, additional evidence are available not for the same student, but rather students of the same racial group. This contrasts with the workplace context in which employers facing the same worker learn about productivity over a longer period and rely progressively less on race as a proxy for ability (Altonji and Pierret, 2001). While we argue that our classroom context does not lend itself to a rigorous examination of learning under statistical discrimination, we acknowledge that accumulated experience could affect teachers' propensity for racially biased evaluations, perhaps by making their in-class assessments more objective (e.g. developing grading rubrics). If so, we should expect that: a) the impact of race and first impressions on teacher evaluation should fall with experience and b) the relation between standardized tests scores and teacher evaluation should strengthen over time. Table OA11 examines this using a sample stratified by teacher experience. While the results do not affirm either element of the hypothesis, the relatively short three year panel may be once again a data limitation that constrains us from drawing definitive conclusions on this point.

Taken together, the evidence suggests some features of teachers' first classrooms do affect sub-

sequent teacher evaluations, although the exact attribute and magnitude vary by context. Teacher assessments are sensitive to the sign and magnitude of the White-Black average test score gap, as well as to the relative position of the lowest Black achiever in the White EOG distribution within initial classrooms. A greater White score advantage and lower achievement for the bottom-scoring Black student (thus higher share of White students who exceed their score) exacerbate racial disparities in assessments when teachers have between 1 and 3 years of experience. Yet we do not observe reduced racial gaps when teachers are exposed to Black academic superstars or when the average Black student outperforms the average White student. The asymmetric responses of novice teachers to the tails of score distributions suggest that they assign greater weights to classroom contexts involving lower-performing students or that these make that first classroom experience more memorable.

6 Conclusion

We use statewide administrative data from North Carolina and document significant racial disparities in teacher assessment. Elementary teachers judge Black students at lower levels of subject mastery than what is indicated by their objectively graded test performance, compared to White peers who are observationally equivalent. These racial differentials hold for both math and reading. In our preferred specification, the conditional racial gap in evaluation is 0.060, which is equivalent to 0.07 standard deviations in the distribution of teacher evaluations. We note an important implication that follows from these findings. Since teacher expectations can shape students' academic achievement and attainment (Papageorge et al., 2016; Hill and Jones, 2017; Lavy and Sand, 2018), we advance that systemic differential assessments unfavorable to specific racial groups can adversely impact student performance. This can lead to under-investment in education for minority groups that in turn perpetuate longstanding achievement gaps. Teacher evaluation differentials can also affect the sorting of students into academic tracks and exacerbate within-school segregation (Clotfelter et al., 2020).

Efforts to bridge gaps can benefit from a more informed understanding of this determinant of achievement disparities.[30] We contribute to the knowledge base by carefully investigating a possible origin of these racial differentials, the early experience of instructors. We hypothesize that novice teachers' initial classroom experiences can leave lasting impressions on the manner in which they assess subsequent students of a given racial or ethnic group. Specifically, we test whether White-Black average score gaps or the existence of high- or low-performing Black students in those initial classrooms influence teachers' subsequent assessment patterns. Having incoming Black students that on average under-performed White students during the teachers' first year of classroom experience reduces their relative evaluations of future cohorts of Black students. Moreover, having an entering Black student who previously scored lower than a larger share of White classmates in that initial class also makes a teacher more likely to underrate Black students relative to their White peers. In contrast, the effects of having a Black student who previously scored *higher* than a larger share of White classmates in that initial class appear muted. Thus existing assessment gaps widen when teachers are exposed to early classrooms conforming to achievement stereotypes, theorized in this context as over-generalized representations of racial group-based differences that allow for more efficient information processing (Hilton and von Hippel, 1996; Bordalo et al., 2016). Notably, the same teachers are not symmetrically updating when exposed to stereotype-defying contexts involving Black superstar students.

Our results are new to the literature in calling attention to the impact of not only exposure to racial groups, but the particular *nature* of those interactions. We show that in the education context, the conditions of early classrooms matter for the formation and reinforcement of racial biases. Importantly, we are not claiming that early experiences explain the entirety of the current racial gaps in assessment. What we find is strong evidence that such first impressions are still relevant up to three years after individuals begin their teaching careers.

These findings that early classroom compositions and racial group-specific performance can shape future assessment practices imply a more deliberate approach to professional development

[30]One particularly promising strategy discussed outside the literature in economics involves the adoption of rubrics for grading, as recently examined in Quinn (2019).

activities and initial classroom assignment. Teachers can be made more aware of the ways in which their early interactions with such students can influence future expectations of minority groups. Alesina et al. (2018), for example, show that revealing implicit association bias test results to teachers is a promising way of combating negative stereotypes towards immigrant children in Italy. Our results call attention to the potential of studying these forms of intervention in combination with controlled allocation of novice teachers to initial classrooms. High volumes of teacher turnover and the growth of teacher training programs like Teach for America continue to boost the population of new teachers, which underscore the urgency of carefully considering the additional consequences of assigning relatively low-performing racial minorities to novice instructors. While most of the education literature has tended to the fact that inexperienced teachers are in general less likely to contribute to learning, our results call attention to an effect that spills over to future cohorts of minority students who interact with teachers in this particular career trajectory.

References

Aigner, D. J. and Cain, G. G. (1977). Statistical Theories of Discrimination in Labor Markets. *Industrial and Labor Relations Review*, 30(2):175–187.

Alesina, A., Carlana, M., Ferrara, E. L., and Pinotti, P. (2018). Revealing Stereotypes: Evidence from Immigrants in Schools. Working Paper 25333, National Bureau of Economic Research.

Altonji, J. G. and Pierret, C. R. (2001). Employer Learning and Statistical Discrimination. *The Quarterly Journal of Economics*, 116(1):313–350.

Ambady, N. and Skowronski, J. J. (2008). *First Impressions*. Guilford Press.

Asch, S. E. (1946). Forming Impressions of Personality. *The Journal of Abnormal and Social Psychology*, 41(3):258–290.

Avitzour, E., Choen, A., Joel, D., and Lavy, V. (2020). On the Origins of Gender-Biased Behavior: The Role of Explicit and Implicit Stereotypes. Technical Report w27818, National Bureau of Economic Research.

Becker, G. S. (1993). *Human capital: a theoretical and empirical analysis, with special reference to education*. The University of Chicago Press, Chicago.

Bond, T. N. and Lang, K. (2013). The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results. *The Review of Economics and Statistics*, 95(5):1468–1479.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Botelho, F., Madeira, R. A., and Rangel, M. A. (2015). Racial Discrimination in Grading: Evidence from Brazil. *American Economic Journal: Applied Economics*, 7(4):37–52.

Burgess, S. and Greaves, E. (2013). Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities. *Journal of Labor Economics*, 31(3):535–576.

Card, D. and Giuliano, L. (2016). Can Tracking Raise the Test Scores of High-Ability Minority Students? *American Economic Review*, 106(10):2783–2816.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9):2633–2679.

Clotfelter, C. T., Ladd, H. F., Clifton, C. R., and Turaeva, M. (2020). School Segregation at the Classroom Level in a Southern 'New Destination' State. *CALDER Working Paper No. 230-0220*.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. (2009). The Academic Achievement Gap in Grades 3 to 8. *The Review of Economics and Statistics*, 91(2):398–419.

Clotfelter, C. T., Ladd, H. F., and Vigdor, J. L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *The Journal of Human Resources*, 41(4):778–820.

Cornell, B. and Welch, I. (1996). Culture, Information, and Screening Discrimination. *Journal of Political Economy*, 104(3):542–571.

Dee, T. S. (2004). Teachers, Race, and Student Achievement in a Randomized Experiment. *The Review of Economics and Statistics*, 86(1):195–210.

Dee, T. S. (2005). A Teacher like Me: Does Race, Ethnicity, or Gender Matter? *The American Economic Review*, 95(2):158–165.

Dee, T. S. (2007). Teachers and the Gender Gaps in Student Achievement. *The Journal of Human Resources*, 42(3):528–554.

DeMeis, D. K. and Turner, R. R. (1978). Effects of Students' Race, Physical Attractiveness, and Dialect on Teachers' Evaluations. *Contemporary Educational Psychology*.

Devine, P. G., Forscher, P. S., Austin, A. J., and Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, 48(6):1267–1278.

Dizon-Ross, R. (2019). Parents' Beliefs about Their Children's Academic Ability: Implications for Educational Investments. *American Economic Review*, 109(8):2728–2765.

Donovan, M. S. and Cross, C. T. (2002). *Minority Students in Special and Gifted Education*. The National Academies Press, Washington, D.C.

Elder, T. and Zhou, Y. (2021). The Black-White Gap in Noncognitive Skills among Elementary School Children. *American Economic Journal: Applied Economics*, 13(1):105–132.

Farkas, G., Grobe, R. P., Sheehan, D., and Shuan, Y. (1990). Cultural Resources and School Success: Gender, Ethnicity, and Poverty Groups within an Urban School District. *American Sociological Review*, 55(1):127–142.

Ferguson, R. F. (1998). Can schools narrow the Black–White test score gap? In *The Black–White test score gap*, pages 318–374. Brookings Institution Press, Washington, DC, US.

Ferguson, R. F. (2003). Teachers' Perceptions and Expectations and the Black-White Test Score Gap. *Urban Education*, 38(4):460–507.

Figlio, D. N. (2005). Names, Expectations and the Black-White Test Score Gap. SSRN Scholarly Paper ID 684721, Social Science Research Network, Rochester, NY.

Fortin, N., Oreopoulos, P., and Phipps, S. (2015). Leaving Boys Behind: Gender Disparities in High Academic Achievement. *Journal of Human Resources*, 50(3):549–579.

Gennaioli, N. and Shleifer, A. (2010). What Comes to Mind. *Quarterly Journal of Economics*, 125(4):1399–1433.

Gershenson, S., Holt, S. B., and Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, 52:209–224.

Hanna, R. N. and Linden, L. L. (2012). Discrimination in Grading. *American Economic Journal: Economic Policy*, 4(4):146–168.

Hedges, L. V. and Nowell, A. (1999). Changes in the Black-White Gap in Achievement Test Scores. *Sociology of Education*, 72(2):111–135.

Hill, A. and Jones, D. B. (2017). Rosenthal Revisited: Self-Fulfilling Prophecies in the Classroom.

Hilton, J. L. and von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, 47:237–271.

Hinnerich, B. T., Höglin, E., and Johannesson, M. (2011). Are boys discriminated in Swedish high schools? *Economics of Education Review*, 30(4):682–690.

Jussim, L. and Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, 9(2):131–155.

Kalogrides, D. and Loeb, S. (2013). Different Teachers, Different Peers: The Magnitude of Student Sorting Within Schools. *Educational Researcher*, 42(6):304–316.

Lang, K. (1986). A Language Theory of Discrimination. *The Quarterly Journal of Economics*, 101(2):363–382.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92(10–11):2083–2105.

Lavy, V. and Megalokonomou, R. (2019). Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study. Technical Report w26021, National Bureau of Economic Research, Cambridge, MA.

Lavy, V. and Sand, E. (2018). On the origins of gender gaps in human capital: Short- and long-term consequences of teachers' biases. *Journal of Public Economics*, 167:263–279.

Leiter, J. and Brown, J. S. (1985). Determinants of Elementary School Grading. *Sociology of Education*, 58(3):166–180.

Lindahl, E. (2016). Are teacher assessments biased? – evidence from Sweden. *Education Economics*, 24(2):224–238.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109.

Lundberg, S. J. and Startz, R. (1983). Private Discrimination and Social Intervention in Competitive Labor Market. *The American Economic Review*, 73(3):340–347.

Macrae, C. N., Stangor, C., and Hewstone, M. (1996). *Stereotypes and Stereotyping*. Guilford Press. Google-Books-ID: o2EVqBMpJDEC.

Mechtenberg, L. (2009). Cheap Talk in the Classroom: How Biased Grading at School Explains Gender Differences in Achievements, Career Choices and Wages. *The Review of Economic Studies*, 76(4):1431–1459.

Natriello, G. and Dornbusch, S. M. (1983). Bringing Behavior Back In: The Effects of Student Characteristics and Behavior on the Classroom Behavior of Teachers. *American Educational Research Journal*, 20(1):29–43.

Neal, D. (2006). Chapter 9 Why Has Black–White Skill Convergence Stopped? In *Handbook of the Economics of Education*, volume 1, pages 511–576. Elsevier.

Nickerson, R. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

Nisbett, R. E. and Ross, L. (1980). *Human inference: strategies and shortcomings of social judgment*. Prentice-Hall.

Papageorge, N. W., Gershenson, S., and Kang, K. (2016). Teacher Expectations Matter. SSRN Scholarly Paper ID 2834215, Social Science Research Network, Rochester, NY.

Papay, J. P., Murnane, R. J., and Willett, J. B. (2016). The Impact of Test Score Labels on Human-Capital Investment Decisions. *Journal of Human Resources*, 51(2):357–388.

Pettigrew, T. F., Tropp, L. R., Wagner, U., and Christ, O. (2011). Recent advances in intergroup contact theory. *International Journal of Intercultural Relations*, 35(3):271–280.

Phillips, M., Crouse, J., and Ralph, J. (1998). Does the Black–White test score gap widen after children enter school? In *The Black–White test score gap*, pages 229–272. Brookings Institution Press, Washington, DC, US.

Quinn, D. M. (2019). Rubrics to Mitigate Racial Bias. *Working paper*, page 56.

Rabin, M. and Schrag, J. L. (1999). First Impressions Matter: A Model of Confirmatory Bias. *The Quarterly Journal of Economics*, 114(1):37–82.

Reardon, S. F. and Galindo, C. (2009). The Hispanic-White Achievement Gap in Math and Reading in the Elementary Grades. *American Educational Research Journal*, 46(3):853–891.

Reardon, S. F. and Robinson, J. P. (2008). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. *Handbook of research in education finance and policy*, pages 497–516.

Rist, R. C. (1973). *The Urban School: A Factory for Failure. A Study of Education in American Society*. MIT Press, Cambridge, Mass.

Rosenthal, R. and Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review*, 3(1):16–20.

Sewell, W. H. and Hauser, R. M. (1980). *The Wisconsin Longitudinal Study of Social and Psychological Factors in Aspirations and Achievements*.*

Sexton, P. C. (1961). *Education and income: inequalities of opportunity in our public schools*. Viking Press.

Sprietsma, M. (2013). Discrimination in grading: experimental evidence from primary school teachers. *Empirical Economics*, 45(1):523–538.

Steele, C. M. and Aronson, J. (1998). Stereotype threat and the test performance of academically successful African Americans. In *The Black–White test score gap*, pages 401–427. Brookings Institution Press, Washington, DC, US.

Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review*, 77:101981.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140.

Williams, T. (1976). Teacher Prophecies and the Inheritance of Inequality. *Sociology of Education*, 49(3):223–236.

Figures and Tables

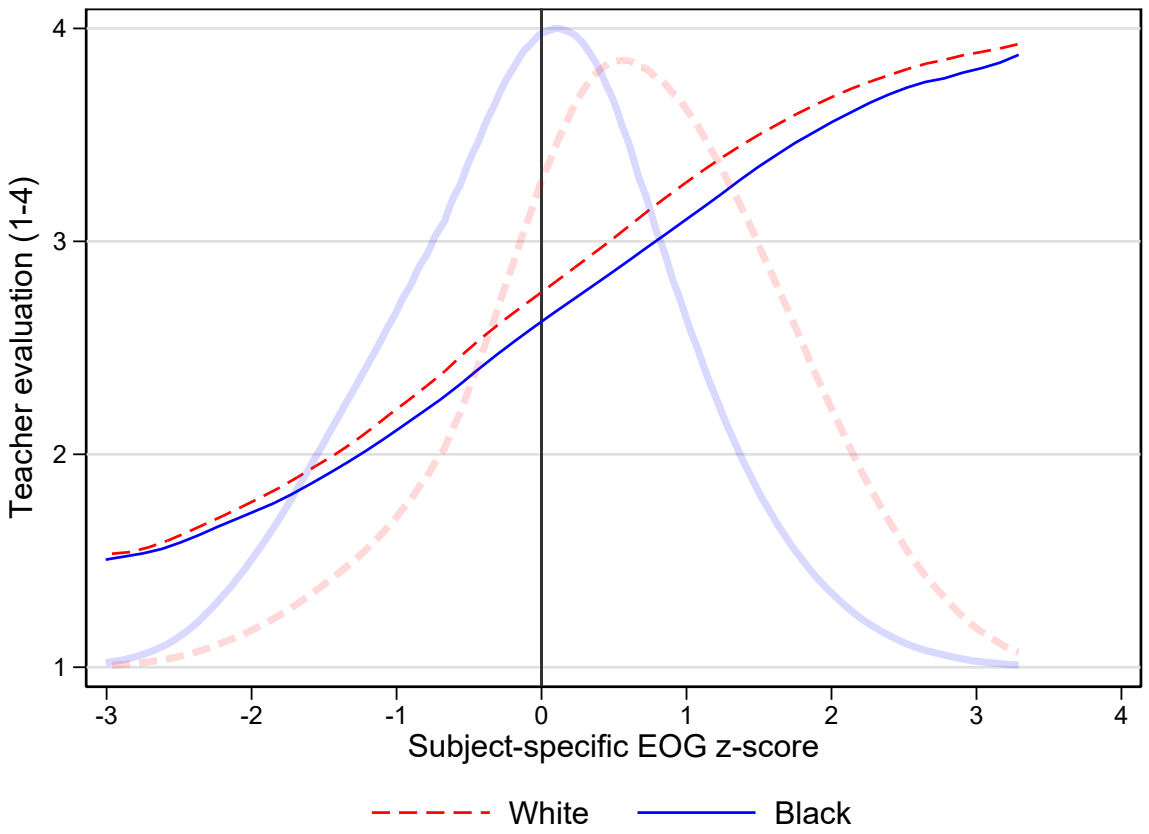
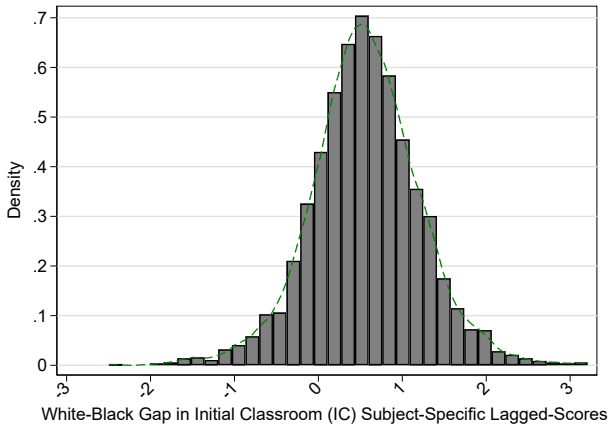
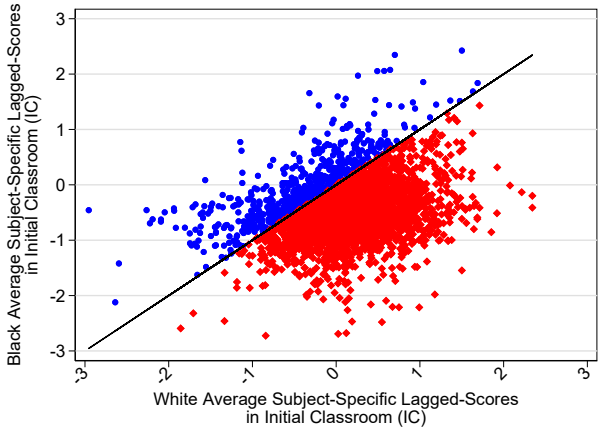


Figure 1: Contemporaneous relation between teacher evaluation and End-of-Grade (EOG) standardized test scores

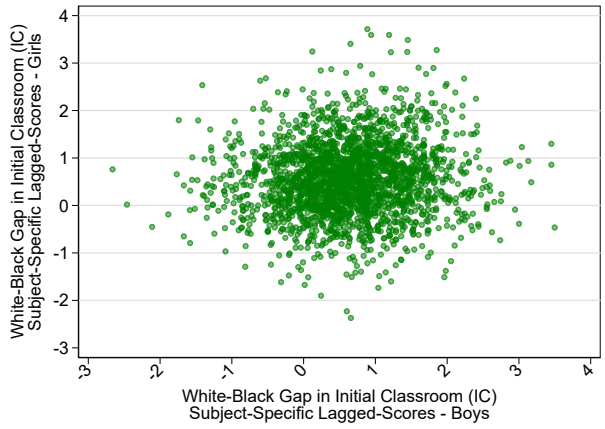
Note: Pooled sample of subjects (math and reading). Sample restricted to novice teachers (1 to 3 years of experience beyond their very first year), which is the working sample for analysis describe in text. Local polynomial estimates of bivariate relationships and race-specific kernel-density estimates for EOG test scores. Teacher evaluation and EOG scores are collected in the Spring of each academic year. While students take the EOG tests, teachers are required to fill a questionnaire which requests their subject-specific subjective evaluation of each student.



Panel A: Distribution of Initial Classroom (IC) White-Black gaps in lagged EOG test scores



Panel B:Initial Classroom (IC) average lagged EOG test scores for Black and White Students



Panel C:Initial Classroom (IC) White-Black gaps in lagged EOG test scores by gender

Figure 2: Initial Classroom (IC) subject-specific lagged end-of-grade (EOG) test scores

Note: Initial-classroom level observations for subset with mixed-race composition. Lagged EOG scores from the previous school year are used to construct IC measures of student performance. These lagged test scores precede any classroom interaction with, and therefore cannot be influenced by, the teachers in our working sample.

Table 1: Descriptive Statistics

	Full Sample		IC Sample	
	Mean	SD	Mean	SD
	(1a)	(1b)	(2a)	(2b)
Panel A: Student-subject-year				
Grade 5	0.52	0.50	0.51	0.50
Math subject	0.50	0.50	0.48	0.50
Age in Dec 31 of EOG test year	10.95	0.76	10.93	0.76
Female	0.50	0.50	0.50	0.50
White	0.47	0.50	0.47	0.50
Black	0.31	0.46	0.31	0.46
Hispanic	0.14	0.35	0.14	0.35
Asian	0.02	0.14	0.02	0.15
American Indian	0.02	0.13	0.01	0.11
Other race or ethnicity	0.04	0.20	0.04	0.20
Days absent	6.03	5.76	6.05	5.74
Chronic absenteeism	0.04	0.19	0.04	0.19
Teacher evaluation (previous school year)	2.72	1.03	2.71	1.03
EOG z-score (centered at proficiency cutoff)	0.33	1.04	0.32	1.04
EOG-achievement level	2.73	0.88	2.74	0.88
Proficient based on EOG score	0.67	0.47	0.67	0.47
Teacher evaluation (current year)	2.85	0.84	2.84	0.84
Proficient based on teacher evaluation	0.70	0.46	0.69	0.46
Above classroom mean based on teacher evaluation	0.52	0.50	0.52	0.50
Observations	203,062		156,291	
Panel B: Classroom-subject-year				
Female teacher	0.88	0.33	0.87	0.33
White teacher	0.91	0.29	0.91	0.28
Black teacher	0.07	0.26	0.07	0.25
Teacher years of experience	1.78	0.79	1.78	0.79
Teacher education: BA only	0.83	0.37	0.84	0.37
Teacher education: MA	0.14	0.34	0.14	0.34
Share of White students in current classroom	0.45	0.25	0.45	0.24
Share of Black students in current classroom	0.32	0.22	0.33	0.21
Number of students	27.26	13.58	26.15	12.61
Grade 5	0.48	0.50	0.47	0.50
Math subject	0.50	0.50	0.49	0.50
Observations	7,494		6,011	
Panel C: Teacher				
Female teacher	0.88	0.32	0.88	0.32
White teacher	0.90	0.30	0.91	0.29
Black teacher	0.08	0.26	0.07	0.25
Teacher education: BA only	0.83	0.38	0.83	0.38
Teacher education: MA	0.14	0.35	0.15	0.35
Teacher licensed in NC	0.56	0.50	0.55	0.50
Share of White students in initial classroom (IC)			0.44	0.24
Share of Black students in initial classroom			0.34	0.21
White-Black score gap in initial classroom			0.55	0.61
Share of White students above highest-scoring Black student in IC			0.33	0.27
Share of White students below lowest-scoring Black student in IC			0.09	0.16
Observations	2,196		1,907	

Notes: For measures at the initial classroom-level, we average across subjects within teachers.

Table 2: Teacher Subject-Mastery Evaluation, Student Race, and Standardized Test Scores (EOG)

	Evaluation Scale (1-4)					Binary Indicators	
	Mean levels	Raw Difference	+Class-Subject FE	+Test scores	+ Demographics	1{Proficient}	1{> Class Mean}
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
White	3.055						
	[$SD = 0.80$]						
Black	2.598	−0.457***	−0.416***	−0.059***	−0.060***	−0.025***	−0.031***
	[$SD = 0.82$]	(0.009)	(0.008)	(0.004)	(0.004)	(0.003)	(0.003)
EOG test scores				0.691***	0.682***	0.375***	0.369***
				(0.004)	(0.004)	(0.003)	(0.003)
Observations	203,062	203,062	203,062	203,062	203,062	203,062	203,062
Classroom-subjects	7,494	7,494	7,494	7,494	7,494	7,494	7,494
Teachers	2,196	2,196	2,196	2,196	2,196	2,196	2,196

Notes: All standard errors are clustered at the teacher's unique ID level. EOG test scores are included as z-scores centered at grade-subject state-mandated proficiency cutoff and as a fourth-order polynomial function. Reported coefficient on EOG test scores is the marginal effect evaluated at the proficiency cutoff. Demographic controls include indicators for gender and age relative to the within-grade modal age (and these are include in all models reported from Columns 5 to 7). Average evaluation score by teachers are 2.85 (SD=0.84) for whole population. Proficiency share is 0.70 (0.58 for Black and 0.79 for White students) and share of evaluated above the classroom mean is 0.52 (0.43 for Black and 0.58 for White students). *** p<0.01, ** p<0.05, * p<0.1

Table 3: Initial Classroom (IC) Conditions and Racial Differentials in Teacher Evaluation (1-4 Scale)

	Full sample	Sample with IC information			
	(1)	(2)	(3)	(4)	(5)
Black	−0.060***	−0.062***	−0.060***	−0.060***	−0.012
	(0.004)	(0.005)	(0.005)	(0.005)	(0.012)
Black × Share of Black in IC			0.028	0.023	0.027
			(0.026)	(0.026)	(0.026)
Black × White-Black score gap in IC				−0.033***	
				(0.009)	
Black × 1{White-Black score gap in IC > 0}					−0.056***
					(0.013)
Observations	203,062	156,291	156,291	156,291	156,291
Classroom-subjects	7,494	6,011	6,011	6,011	6,011
Teachers	2,196	1,907	1,907	1,907	1,907

Notes: All models are estimated using the set of controls listed in Col 5 of Table 2, which include EOG test scores (fourth-order polynomial), gender, and age indicators. The initial classroom (IC) sample restricts to observations with racial mix (at least one student of each race) and measured White-Black gaps in lagged test scores. Shares of Black students and the White-Black score gap in initial classrooms are centered at sample means for interactions. All interactions with other non-White races/ethnicities are also included in the model so that coefficients juxtaposes between Black and White students. *** p<0.01, ** p<0.05, * p<0.1

Table 4: Initial Classroom (IC) Conditions, Future Classroom (FC) Conditions and Racial Differentials in Teacher Evaluation (1-4 Scale)

	Initial Classroom (IC) Conditions			Future Classroom (FC) Conditions		
	(1)	(2)	(3)	(4)	(5)	(6)
Black	−0.058***	−0.058***	−0.004	−0.059***	−0.059***	−0.069***
	(0.010)	(0.010)	(0.021)	(0.010)	(0.010)	(0.017)
Black × Share of Black in IC or FC	0.023	0.020	0.027	0.007	0.005	0.008
	(0.049)	(0.048)	(0.049)	(0.049)	(0.049)	(0.049)
Black × White-Black score gap in IC or FC		−0.028*			−0.009	
		(0.016)			(0.015)	
Black × 1{White-Black score gap in IC>0 or FC>0}			−0.063***			0.013
			(0.024)			(0.019)
Observations	40,051	40,051	40,051	40,051	40,051	40,051
Classroom-subjects	1,597	1,597	1,597	1,597	1,597	1,597
Teachers	389	389	389	389	389	389

Notes: All models are estimated using the set of controls listed in Col 5 of Table 2, which include EOG test scores (fourth-order polynomial), and gender and age indicators. The initial/future classroom sample restricts to observations with racial composition information (at least one student of each race). Shares of Black students and the White-Black score gap in initial/future classrooms are centered at sample means for interactions. All interactions with other non-White races/ethnicities are also included in the model so that coefficients juxtaposes between Black and White students. *** p<0.01, ** p<0.05, * p<0.1

Table 5: Strata by Sign of White-Black Score Gap in Initial (IC) or Future Classroom (FC)

	White-Black Gap in IC$\leqslant 0$ or in FC$\leqslant 0$ (1)	White-Black Gap in IC> 0 or in FC> 0 (2)	Difference (3)=(2)-(1)
Panel A: Initial Classroom (IC) conditions			
Black	-0.021^{*}	-0.068^{***}	-0.048^{***}
	(0.011)	(0.005)	(0.012)
Male	-0.044^{***}	-0.035^{***}	0.009
	(0.008)	(0.004)	(0.009)
Observations	24,639	131,652	156,291
Panel B: Initial Classroom (IC) conditions, sub-sample with measured FC Conditions			
Black	-0.016	-0.065^{***}	-0.049^{**}
	(0.022)	(0.011)	(0.024)
Male	-0.051^{***}	-0.043^{***}	0.008
	(0.015)	(0.007)	(0.017)
Observations	6,005	34,040	40,045
Panel C: Future Classroom (FC) conditions			
Black	-0.061^{***}	-0.058^{***}	0.003
	(0.018)	(0.011)	(0.021)
Male	-0.051^{***}	-0.042^{***}	0.009
	(0.016)	(0.007)	(0.017)
Observations	6,998	33,047	40,045
Panel D: Replace IC's White-Black score gap with IC's Female-Male score gap			
Black	-0.057^{***}	-0.066^{***}	-0.008
	(0.007)	(0.007)	(0.009)
Male	-0.029^{***}	-0.044^{***}	-0.015^{**}
	(0.005)	(0.005)	(0.006)
Observations	75,888	80,403	156,291

Notes: All models are estimated using the set of controls listed in Col 5 of Table 2, which include EOG test scores (fourth-order polynomial), gender, and age indicators. The initial classroom/future sample restricts to observations with racial mix (at least one student of each race) and measured White-Black gaps in lagged test scores. Shares of Black students are centered at sample means for interactions used as controls in these models. All interactions with other non-White races/ethnicities are also included in the model so that coefficients juxtaposes between Black and White students. *** p<0.01, ** p<0.05, * p<0.1

Table 6: Initial Classroom (IC) Conditions and Racial Differentials in Teacher Evaluation (1-4 Scale) – strata by current student gender and accounting for gender-specific IC racial gap

	Boys only		Girls only	
	(1)	(2)	(3)	(4)
Black	−0.061***	−0.064***	−0.072***	−0.075***
	(0.008)	(0.008)	(0.008)	(0.008)
Black × Share of Black in IC	−0.020	−0.020	0.069	0.069
	(0.046)	(0.046)	(0.048)	(0.048)
Black × White-Black score gap in IC	−0.040**		−0.035**	
	(0.016)		(0.015)	
Black × White Boy-Black Boy score gap in IC		−0.027**		−0.002
		(0.011)		(0.010)
Black × White Girl-Black Girl score gap in IC		−0.009		−0.034***
		(0.011)		(0.011)
Observations	51,037	51,037	49,581	49,581
Classroom-subjects	3,854	3,854	3,737	3,737
Teachers	1,251	1,251	1,219	1,219

Notes: All models are estimated using the set of controls listed in Col 5 of Table 2, which include EOG test scores (fourth-order polynomial), gender, and age indicators. The initial classroom (IC) sample restricts to observations with racial mix (at least one student of each race) and measured White-Black gaps in lagged test scores. Shares of Black students and the White-Black score gap in initial classrooms are centered at sample means for interactions. All interactions with other non-White races/ethnicities are also included in the model so that coefficients juxtaposes between Black and White students. *** p<0.01, ** p<0.05, * p<0.1

Table 7: Tails of the Initial Classroom (IC) Performance Distribution and Teacher Evaluation Bias

	(1)	(2)	(3)	(4)	(5)	(6)
Black × Pr[White outscores highest-scoring Black student in IC]	−0.033*		−0.020	−0.011		0.000
	(0.020)		(0.020)	(0.020)		(0.020)
Black × Pr[White outscores lowest-scoring Black student in IC]		−0.109***	−0.102***		−0.099***	−0.100***
		(0.028)	(0.029)		(0.033)	(0.033)
Interacted controls						
Black × Female teacher	−0.023	−0.024	−0.024	−0.012	−0.012	−0.012
	(0.016)	(0.016)	(0.016)	(0.017)	(0.017)	(0.017)
Black × Black teacher	0.013	0.014	0.014	0.003	0.005	0.005
	(0.027)	(0.027)	(0.027)	(0.025)	(0.025)	(0.025)
Black × Teacher has MA	−0.054***	−0.054***	−0.054***	−0.057***	−0.057***	−0.057***
	(0.016)	(0.016)	(0.016)	(0.018)	(0.018)	(0.018)
Black × Teacher licensed in NC	−0.026**	−0.028***	−0.028**	−0.026**	−0.028**	−0.028**
	(0.011)	(0.011)	(0.011)	(0.013)	(0.013)	(0.013)
Black × Teacher experience	−0.007	−0.007	−0.007	−0.007	−0.007	−0.007
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Black × initial school FE	NO	NO	NO	YES	YES	YES
Observations	156,291	156,291	156,291	156,285	156,285	156,285

Notes: All models are estimated using the set of controls listed in Col 5 of Table 3, which include EOG test scores, gender, age and month of birth indicators. The initial classroom sample restricts to observations with racial composition information. Share of Black students in initial classrooms is centered at sample mean for interactions. All interactions with other non-White races and ethnicities are also included in the model so that coefficients juxtaposes between Black and White students. Models in columns 4 to 6 exclude singletons. *** p<0.01, ** p<0.05, * p<0.1.

Online Appendix

for Rangel and Shi's "First Impressions Matter"

A. Tables and Figures

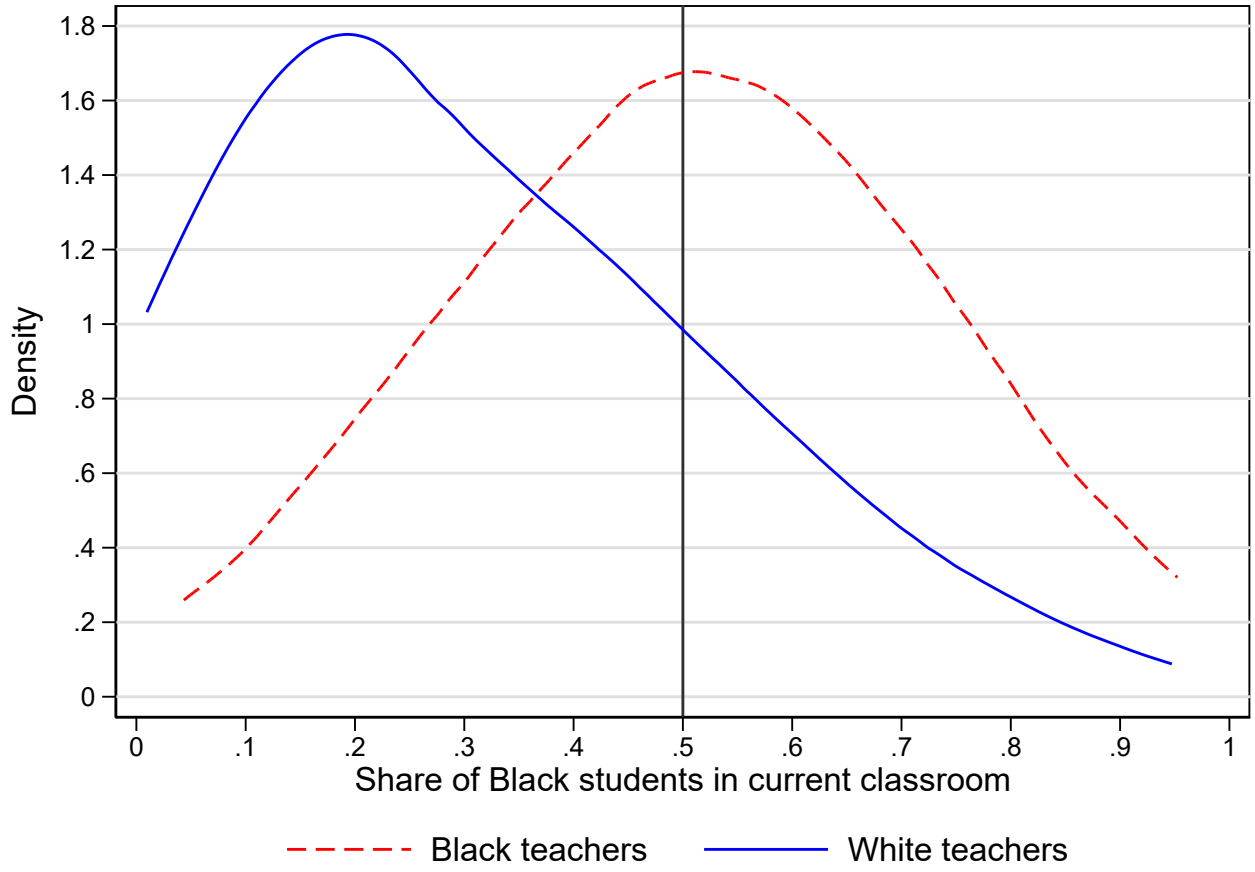
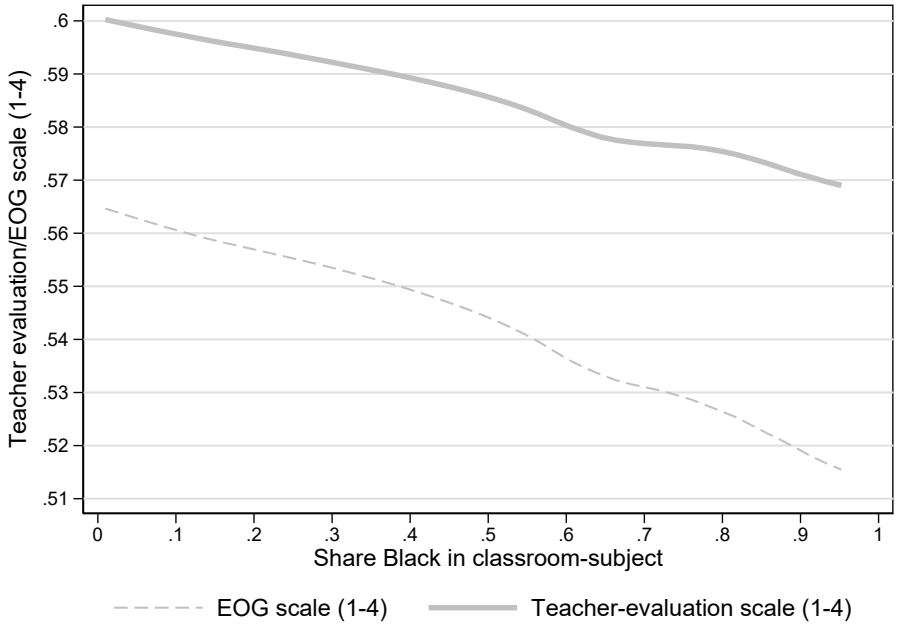
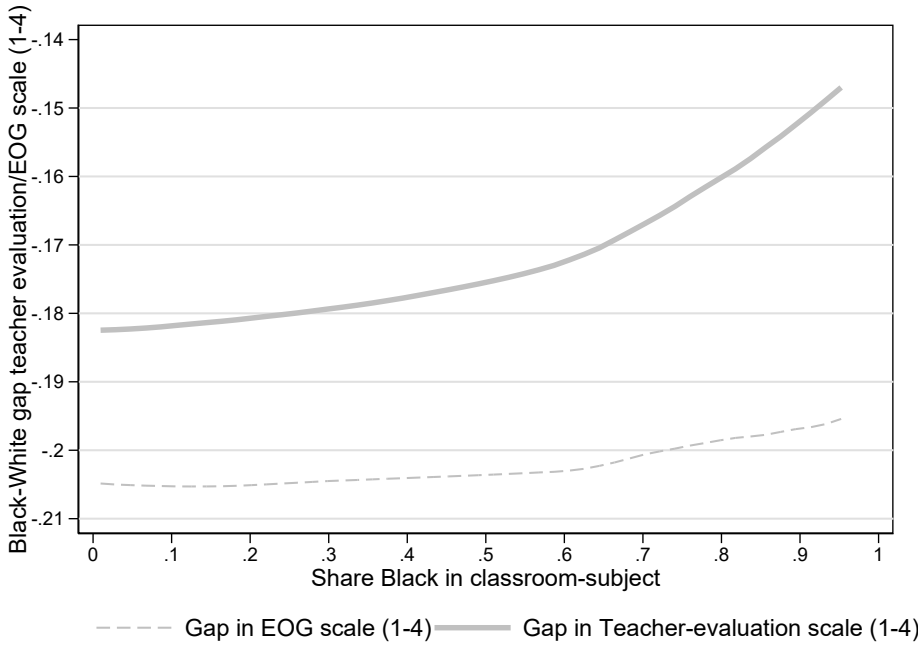


Figure OA1: Classroom-level racial composition by teacher race



Panel A: Black Students' Teacher Evaluation and Standardized Test Score levels



Panel B: Black-White Gap in Teacher Evaluation and Standardized Test Score levels

Figure OA2: Classroom-level objective and subjective measures versus racial composition

Table OA1: Variance decomposition within and across classrooms

	Within variation share (1)	Between variation share (2)
Panel A: Our working sample		
a. Teacher evaluation levels (1-4) - Math	0.86	0.14
b. EOG achievement levels (1-4) - Math	0.74	0.26
Ratio a/b	1.16	0.54
Difference a-b	0.12	-0.12
c. Teacher evaluation levels (1-4) - Reading	0.86	0.14
d. EOG achievement levels (1-4) - Reading	0.79	0.21
Ratio c/d	1.09	0.67
Difference c-d	0.07	-0.07
Observations Math	101,995	101,995
Observations Reading	101,067	101,067
Panel B: Elder and Zhou (2020) - 3rd graders, ECLS-K 2011 [Table 4, column 4]		
a. Math rating	0.99	0.08
b. IRT test scores - Math	0.69	0.31
Ratio a/b	1.44	0.26
Difference a-b	0.30	-0.30
c. Reading rating	0.91	0.09
d. IRT test scores - Reading	0.68	0.32
Ratio c/d	1.34	0.28
Difference c-d	0.23	-0.23
Observations	NA	NA

Table OA2: Initial Classroom Characteristics and Novice Teacher Attributes - Novice-teacher allocation

	Share Black (1)	Average EOG, Black (2)	White-Black Gap (3)
Panel A: Raw analysis			
Female teacher	−0.009	−0.026	0.023
	(0.015)	(0.032)	(0.044)
Black teacher	0.160***	−0.075*	0.075
	(0.018)	(0.040)	(0.058)
Teacher education: BA only	0.038***	−0.067*	−0.100**
	(0.014)	(0.034)	(0.040)
Teacher Licensed	−0.045***	0.036	0.022
	(0.010)	(0.024)	(0.029)
Observations	1,907	1,907	1,907
Panel B: Conditional on initial school fixed effects			
Female teacher	−0.003	0.011	0.006
	(0.009)	(0.037)	(0.054)
Black teacher	0.001	−0.051	0.130*
	(0.012)	(0.047)	(0.070)
Teacher education: BA only	0.006	0.019	−0.010
	(0.008)	(0.043)	(0.051)
Teacher Licensed	0.003	0.022	−0.011
	(0.006)	(0.030)	(0.039)
Observations	1,575	1,575	1,575

Notes: This table shows the correspondence between novice teacher observable characteristics and those of their initial classrooms. Sample after school-fixed effects is reduced due to singletons. *** p<0.01, ** p<0.05, * p<0.1.

Table OA3: Robustness to Alternative Specifications of EOG Test Scores

	Common support Trimmed tails	Common support Discretized	EOG FE
	(1)	(2)	(3)
Panel A: Teacher Evaluation Scale (1-4)			
Black	−0.060***	−0.057***	−0.057***
	(0.004)	(0.007)	(0.004)
EOG test score	0.682***	0.683***	
	(0.004)	(0.013)	
Panel B: Teacher Evaluation (Proficient=1)			
Black	−0.024***	−0.022***	−0.023***
	(0.003)	(0.004)	(0.003)
EOG test score	0.375***	0.391***	
	(0.003)	(0.009)	
Panel C: Teacher Evaluation (Above class mean=1)			
Black	−0.031***	−0.034***	−0.028***
	(0.003)	(0.004)	(0.003)
EOG test score	0.369***	0.394***	
	(0.003)	(0.008)	
Observations	187,536	39,422	203,042

Notes: This table uses alternative specifications of EOG test scores to examine whether there is sufficient within-classroom overlap of Black and White test score distributions. Column 1 trims the tails of performance distributions so that the sample ranges from the maximum of the lowest Black and White scorers up to the minimum of the top scorers by race. Column 2 only keeps observations for which a Black student has a White classmate with the same EOG score for a given subject. Column 3 includes EOG fixed effects. All standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1.

Table OA4: Robustness of Teacher Evaluation Bias to Inclusion of Socio-demographic, Academic, and Behavioral Covariates – Teacher Evaluation Scale (1-4)

	(1)	(2)	(3)	(4)	(5)	(6)
Black	−0.060***	−0.053***	−0.071***	−0.065***	−0.059***	−0.054***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.006)	(0.006)
EOG test scores	0.682***	0.544***	0.682***	0.679***	0.687***	0.657***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.006)	(0.006)
Male	−0.036***	−0.028***	−0.035***	−0.029***	−0.031***	−0.025***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
One year younger than mode	0.038	0.028	0.030	0.031	0.040	0.035
	(0.039)	(0.039)	(0.039)	(0.040)	(0.047)	(0.048)
One year older than mode	−0.052***	−0.037***	−0.048***	−0.047***	−0.046***	−0.045***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Two years older than mode	−0.164***	−0.111***	−0.150***	−0.146***	−0.145***	−0.140***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.012)	(0.012)
Three+ years older than mode	−0.163***	−0.112*	−0.138**	−0.136**	−0.120	−0.109
	(0.057)	(0.057)	(0.058)	(0.058)	(0.073)	(0.072)
Additional controls included						
Days absent indicators		YES	YES	YES	YES	YES
Lag student evaluation indicators		YES				
Suspension days indicators				YES	YES	YES
Free-reading indicators						YES
Homework indicators						YES
Sample restrictions (due to information availability)						
Restricted to 2008-2013			YES	YES		
Restricted to 2008-2012					YES	YES
Observations	203,062	203,062	195,471	195,471	118,579	118,579

Notes: All standard errors are clustered at the teacher level. EOG test scores are included as z-scores centered at state-mandated proficiency cutoff and as a fourth-order polynomial function. Reported coefficient on EOG test scores is the marginal effect evaluated at the proficiency cutoff. Additional controls are accounted for semi-parametrically with a set of indicator functions. Days of suspension are included separately for each infraction. *** p<0.01, ** p<0.05, * p<0.1

Table OA5: Teacher Subject-Mastery Evaluation Scale (1-4), Student Race, and Standardized Test Scores (EOG) - by subject/grade/teacher experience strata

	By subject		By grade		By teacher experience		
	Math	Reading	Grade 4	Grade 5	1 year	2 years	3 years
	[1]	[2]	[3]	[4]	[5]	[6]	[7]
Black	−0.039***	−0.079***	−0.055***	−0.063***	−0.062***	−0.057***	−0.058***
	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.008)	(0.009)
EOG test score	0.714***	0.657***	0.701***	0.666***	0.674***	0.682***	0.698***
	(0.006)	(0.005)	(0.006)	(0.006)	(0.006)	(0.007)	(0.009)
Male	−0.026***	−0.051***	−0.043***	−0.031***	−0.033***	−0.050***	−0.024***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.006)
One year younger than mode	0.017	0.062	0.074	0.008	0.019	0.113*	−0.025
	(0.045)	(0.047)	(0.063)	(0.048)	(0.058)	(0.058)	(0.091)
One year older than mode	−0.036***	−0.065***	−0.050***	−0.054***	−0.058***	−0.050***	−0.045***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
Two years older than mode	−0.131***	−0.193***	−0.165***	−0.163***	−0.167***	−0.155***	−0.171***
	(0.011)	(0.011)	(0.014)	(0.012)	(0.014)	(0.015)	(0.017)
Three+ years older than mode	−0.091	−0.233***	−0.243***	−0.067	−0.249***	−0.180*	0.012
	(0.064)	(0.068)	(0.066)	(0.095)	(0.069)	(0.092)	(0.143)
Observations	101,995	101,067	96,831	106,231	87,749	67,419	47,894
Classroom-subjects	3,750	3,744	3,928	3,566	3,332	2,444	1,718
Teachers	2,195	2,191	1,210	1,160	1,644	1,207	846

Notes: All standard errors are clustered at the teacher's unique ID level. EOG test scores are included as z-scores centered at grade-subject state-mandated proficiency cutoff and as a fourth-order polynomial function. Reported coefficient on EOG test scores is the marginal effect evaluated at the proficiency cutoff. Demographic controls include indicators for gender and age relative to the within-grade modal age. *** p<0.01, ** p<0.05, * p<0.1

Table OA6: Teacher Subject-Mastery Evaluation Scale (1-4), Student Race, and Standardized Test Scores (EOG) - by child gender and age strata

	By gender		By age	
	Girls	Boys	At or below mode	Above mode
	[1]	[2]	[3]	[4]
Black	−0.068***	−0.054***	−0.067***	−0.044***
	(0.006)	(0.006)	(0.005)	(0.007)
EOG test score	0.674***	0.683***	0.673***	0.696***
	(0.006)	(0.005)	(0.005)	(0.006)
Male	-	-	−0.032***	−0.045***
			(0.004)	(0.006)
One year younger than mode	−0.025	0.110	0.031	-
	(0.054)	(0.068)	(0.041)	
One year older than mode	−0.046***	−0.058***	-	0.120**
	(0.005)	(0.005)		(0.061)
Two years older than mode	−0.133***	−0.184***	-	0.017
	(0.014)	(0.012)		(0.061)
Three+ years older than mode	−0.082	−0.221***	-	-
	(0.099)	(0.077)		
Observations	89,170	90,828	114,513	64,984
Classroom-subjects	6,313	6,520	6,688	5,774
Teachers	1,943	1,987	2,025	1,851

Notes: All standard errors are clustered at the teacher's unique ID level. EOG test scores are included as z-scores centered at grade-subject state-mandated proficiency cutoff and as a fourth-order polynomial function. Reported coefficient on EOG test scores is the marginal effect evaluated at the proficiency cutoff. Demographic controls include indicators for gender and age relative to the within-grade modal age. *** p<0.01, ** p<0.05, * p<0.1

Table OA7: Initial Classroom (IC) Conditions and Racial Differentials in Teacher Evaluation (1-4 Scale) – strata by current student gender

	Full sample	Sample with IC information			
	(1)	(2)	(3)	(4)	(5)
Panel A: Boys only					
Black	−0.054***	−0.055***	−0.053***	−0.052***	−0.006
	(0.006)	(0.007)	(0.007)	(0.007)	(0.016)
Black × Share of Black in IC			0.013	0.008	0.011
			(0.035)	(0.035)	(0.035)
Black × White-Black score gap in IC				−0.030***	
				(0.011)	
Black × 1{White-Black score gap in IC > 0}					−0.054***
					(0.017)
Observations	90,828	71,259	71,259	71,259	71,259
Classroom-subjects	6,520	5,345	5,345	5,345	5,345
Teachers	1,987	1,770	1,770	1,770	1,770
Panel B: Girls only					
Black	−0.068***	−0.071***	−0.071***	−0.070***	−0.012
	(0.006)	(0.007)	(0.007)	(0.007)	(0.016)
Black × Share of Black in IC			0.054	0.050	0.055
			(0.036)	(0.035)	(0.036)
Black × White-Black score gap in IC				−0.039***	
				(0.012)	
Black × 1{White-Black score gap in IC > 0}					−0.069***
					(0.017)
Observations	89,170	69,515	69,515	69,515	69,515
Classroom-subjects	6,313	5,164	5,164	5,164	5,164
Teachers	1,943	1,726	1,726	1,726	1,726

Notes: All models are estimated using the set of controls listed in Col 5 of Table 2, which include EOG test scores (fourth-order polynomial), gender, and age indicators. The initial classroom (IC) sample restricts to observations with racial mix (at least one student of each race) and measured White-Black gaps in lagged test scores. Shares of Black students and the White-Black score gap in initial classrooms are centered at sample means for interactions. All interactions with other non-White races/ethnicities are also included in the model so that coefficients juxtaposes between Black and White students. *** p<0.01, ** p<0.05, * p<0.1

Table OA8: Robustness of Effect of Initial Classroom (IC) Conditions to Interactions with Teacher Attributes

	Base model (1)	+Black × covariates (2)	+Black × initial school FE (3)
Panel A: White-Black score gap in IC			
Black × White-Black score gap in IC	−0.033***	−0.032***	−0.020**
	(0.009)	(0.009)	(0.009)
Interacted controls			
Black × Female teacher		−0.022	−0.011
		(0.016)	(0.017)
Black × Black teacher		0.015	0.005
		(0.027)	(0.025)
Black × Teacher has MA		−0.052***	−0.057***
		(0.016)	(0.018)
Black × Teacher licensed in NC		−0.026**	−0.027**
		(0.011)	(0.013)
Black × Teacher experience		−0.007	−0.007
		(0.006)	(0.006)
Panel B: Sign of White-Black score gap in IC			
Black × 1{White-Black score gap in IC > 0}	−0.056***	−0.054***	−0.039***
	(0.013)	(0.013)	(0.013)
Interacted controls			
Black × Female teacher		−0.021	−0.011
		(0.016)	(0.017)
Black × Black teacher		0.015	0.005
		(0.027)	(0.025)
Black × Teacher has MA		−0.052***	−0.056***
		(0.016)	(0.018)
Black × Teacher licensed in NC		−0.027**	−0.027**
		(0.011)	(0.013)
Black × Teacher experience		−0.006	−0.007
		(0.006)	(0.006)
Observations	156,291	156,291	156,285

Notes: All models are estimated using the set of controls listed in Col 5 and 6 of Table 3, which include EOG test scores, gender, age and month of birth indicators, as well as interactions of Black and IC share of Black students. The initial classroom sample restricts to observations with racial composition information. Shares of Black students and the White-Black score gap in initial classrooms are centered at sample means for interactions. All interactions with other non-White races and ethnicities are also included in the model so that coefficients juxtaposes between Black and White students. Observations in Column 3 exclude singletons. *** p<0.01, ** p<0.05, * p<0.1.

Table OA9: Top and Bottom of the Initial Classroom (IC) Performance Distribution and Teacher Evaluation Bias

	(1)	(2)	(3)	(4)	(5)	(6)
Black × 1{Highest-scoring White student outscores highest-scoring Black student in IC}	−0.003		−0.001	0.003		0.003
	(0.013)		(0.013)	(0.013)		(0.013)
Black × 1{Lowest-scoring White student outscores lowest-scoring Black student in IC}		−0.036***	−0.036***		−0.032***	−0.032***
		(0.012)	(0.012)		(0.012)	(0.012)
Interacted controls						
Black × Female teacher	−0.023	−0.024	−0.024	−0.012	−0.012	−0.012
	(0.016)	(0.016)	(0.016)	(0.017)	(0.017)	(0.017)
Black × Black teacher	0.013	0.012	0.012	0.003	0.004	0.004
	(0.027)	(0.028)	(0.028)	(0.025)	(0.025)	(0.025)
Black × Teacher has MA	−0.055***	−0.054***	−0.054***	−0.057***	−0.056***	−0.056***
	(0.016)	(0.016)	(0.016)	(0.018)	(0.018)	(0.018)
Black × Teacher licensed in NC	−0.027**	−0.029***	−0.029***	−0.026**	−0.028**	−0.028**
	(0.011)	(0.011)	(0.011)	(0.013)	(0.013)	(0.013)
Black × Teacher experience	−0.007	−0.006	−0.007	−0.007	−0.007	−0.007
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Black × initial school FE	NO	NO	NO	YES	YES	YES
Observations	156,291	156,291	156,291	156,285	156,285	156,285

Notes: All models are estimated using the set of controls listed in Col 5 of Table 3, which include EOG test scores, gender, age and month of birth indicators. The initial classroom sample restricts to observations with racial composition information. Share of Black students in initial classrooms is centered at sample mean for interactions. All interactions with other non-White races and ethnicities are also included in the model so that coefficients juxtaposes between Black and White students. Models in columns 4 to 6 exclude singletons. *** p<0.01, ** p<0.05, * p<0.1.

Table OA10: Initial (IC) and Second Classroom (SC) Conditions and Racial Differentials in Teacher Evaluation (1-4 Scale)

	(1)	(2)	(3)	(4)
Black	-0.062^{***}	-0.060^{***}	-0.062^{***}	-0.062^{***}
	(0.012)	(0.012)	(0.012)	(0.012)
Black $\times$ White-Black score gap in IC		-0.044^{**}		-0.041^{*}
		(0.021)		(0.021)
Black $\times$ White-Black score gap in SC			-0.032	-0.022
			(0.022)	(0.023)
Observations	22,680	22,680	22,680	22,680
Classroom-subjects	829	829	829	829
Teachers	445	445	445	445

Notes: All models are estimated using the set of controls listed in Col 5 of Table 2, which include EOG test scores (fourth-order polynomial), gender, and age indicators. The sample restricts to observations with racial mix (at least one student of each race) and measured White-Black gaps in lagged test scores for both IC and SC. Shares of Black students and the White-Black score gap in initial classrooms are centered at sample means for interactions. Classroom racial compositions (IC and SC) are controlled for on the models in Columns 2 to 4. All interactions with other non-White races/ethnicities are also included in the model so that coefficients juxtaposes between Black and White students. *** p<0.01, ** p<0.05, * p<0.1

Table OA11: Initial Classroom (IC) Conditions and Racial Differentials in Teacher Evaluation (1-4 Scale) - stratified by teacher experience

	Full sample	Sample with IC information			
	(1)	(2)	(3)	(4)	(5)
Panel A: teachers with 1 year of experience					
Black	−0.062***	−0.060***	−0.056***	−0.056***	−0.031**
	(0.006)	(0.007)	(0.007)	(0.007)	(0.016)
EOG test scores	0.674***	0.673***	0.673***	0.673***	0.673***
	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Black × Share of Black in IC			0.057	0.052	0.056
			(0.037)	(0.037)	(0.037)
Black × White-Black score gap in IC				−0.027**	
				(0.012)	
Black × 1{White-Black score gap in IC > 0}					−0.030*
					(0.017)
Observations	87749	67485	67485	67485	67485
Panel B: teachers with 2 years of experience					
Black	−0.057***	−0.064***	−0.064***	−0.065***	0.012
	(0.008)	(0.008)	(0.009)	(0.009)	(0.022)
EOG test scores	0.682***	0.678***	0.678***	0.677***	0.677***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Black × Share of Black in IC			0.020	0.017	0.020
			(0.043)	(0.043)	(0.043)
Black × White-Black score gap in IC				−0.027*	
				(0.015)	
Black × 1{White-Black score gap in IC > 0}					−0.090***
					(0.024)
Observations	67419	51782	51782	51782	51782
Panel C: teachers with 3 years of experience					
Black	−0.058***	−0.062***	−0.060***	−0.060***	−0.007
	(0.009)	(0.010)	(0.010)	(0.010)	(0.024)
EOG test scores	0.698***	0.696***	0.696***	0.696***	0.696***
	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)
Black × Share of Black in IC			−0.015	−0.020	−0.014
			(0.046)	(0.046)	(0.046)
Black × White-Black score gap in IC				−0.052***	
				(0.016)	
Black × 1{White-Black score gap in IC > 0}					−0.061**
					(0.026)
Observations	47894	37024	37024	37024	37024

Notes: See Notes in Table 3. *** p<0.01, ** p<0.05, * p<0.1

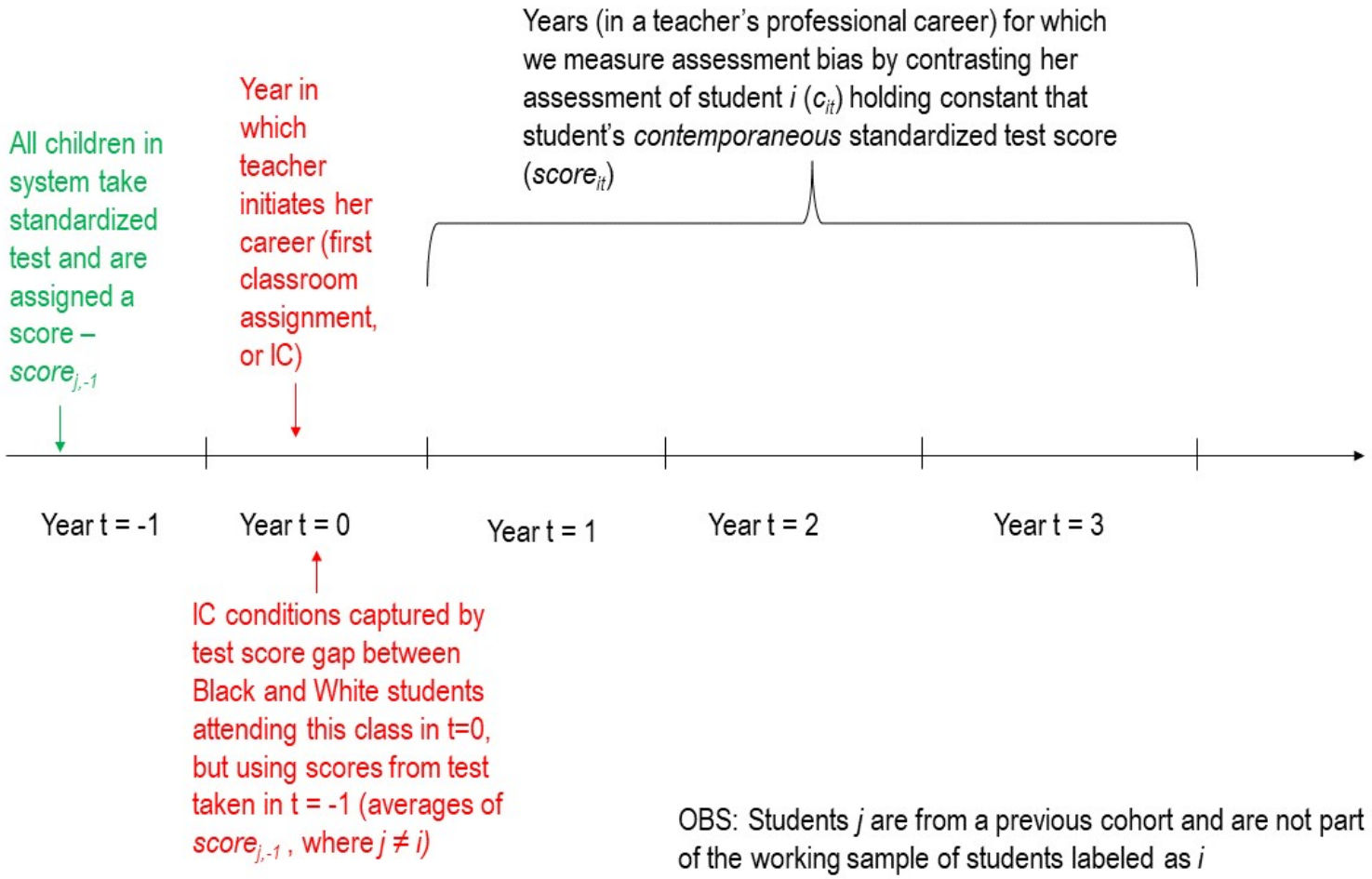


Diagram OA1: Data Organization

B. More on reference bias

We consider the possibility that teachers may rely on relative comparisons across students for their subjective assessments. In fact, it is precisely because of this concern that our estimates of racial differences rely on within-classroom variation and are anchored on blind-scored standardized tests. Based on our reading of Elder and Zhou (2021) and additional examinations of our data presented below, we conclude that our empirical approach sufficiently approximates their strategies to correct for reference bias. We first emphasize that within our study context, there are reasons to believe relative comparisons are accompanied by notions of absolute performance clearly communicated to teachers by education authorities. The North Carolina Department of Public Instruction collects teacher evaluation data as part of the regular testing procedure (distinct from the Early Childhood Longitudinal Study IES-sponsored survey questionnaires utilized in Elder and Zhou's analyses). As we pointed out in the text, teachers were told explicitly to focus their evaluation of student mastery only on the tested subject in state-administered exams and were given a scale that is identical to the one used for standardized exams.

To complement this view, we present additional evidence that our context seems less prone to concerns about reference bias relative to the one faced by Elder and Zhou (2021) after undertaking some of the same analyses. Variance decompositions in Table OA1 indicate that differences in the contribution of between-classroom variation in subjective and objective measures of performance are closer to each other in our sample than in Elder and Zhou (2021), Table 4. Figure OA1 then shows that classroom racial composition and subjective and objective measures of performance follow each other more closely in our study context than in Elder and Zhou (2021), Figure 1, Panels A, B, D and E. While both share a negative slope in our sample, Elder and Zhou (2021) report a positive slope between teacher evaluations and school-level share of Black students, while showing a negative slope between objective test scores and the share of Black students. Most importantly, we show that the Black-White gap in both subjective and objective evaluations in our sample follow a parallel pattern across different classroom compositions. These are indications that our sample does not suffer substantially from the potential reference biases raised by the authors in the context of ECLS-K.

Ultimately, Elder and Zhou (2021) derive Black-White gaps in non-cognitive skills that address reference bias by assuming that variation in objective measures such as tests of cognitive skill are informative about latent distributions in non-cognitive skills. In their second approach, they generate a measure of school-level reference bias by taking the difference between the average objective, or blind-scored, measure of cognitive skill and the average subjective measure of cognitive skill. In doing so they make an assumption that we share in our paper, which is that blind-graded standardized test scores are free of reference bias. Deviations in subjective teacher ratings from these objective measures of cognitive skills at the school level are then added as school-specific reference bias to observed non-cognitive skills to get an adjusted Black-White non-cognitive gap.

Finally, we differ in an important way from Elder and Zhou (2021). Instead of relying on within-school variation (a design that also underpins their third suggested approach), we use classroom fixed effects. This means our racial gaps rely on the very localized context of within-class variation alone, and net out any classroom-specific biases such as the propensity to give everyone uniformly higher ratings. This furthermore includes class-specific reference bias shared across students.