ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-228

Teacher Biases: Evidence from Early-Career Experiences

Marcos A. Rangel
Duke University

Ying Shi
Syracuse University

We study racial bias in teacher assessments using novel statewide data on blind and non-blind evaluations of fourth and fifth graders. Teachers are both significantly less likely to overrate and more likely to under-rate black students' math and reading skills relative to their white classmates. We show that biased evaluations are significantly influenced by initial classroom experiences. Teachers who begin their careers in classrooms with large black-white score gaps carry negative views into evaluations of future cohorts of black students, being particularly sensitive to the lowest-performing black students in early classrooms. This is consistent with the operation of confirmatory biases.

VERSION: June 2020

Suggested citation: Rangel, Marcos A., and Ying Shi. (2020). Teacher Biases: Evidence from Early-Career Experiences. (EdWorkingPaper: 20-228). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/gg76-dv20

FIRST IMPRESSIONS: THE CASE OF TEACHER RACIAL BIAS

MARCOS A. RANGEL
Duke University*

YING SHI
Syracuse University†

June 2020

Abstract

We study racial bias and the persistence of first impressions in the context of education. Teachers who begin their careers in classrooms with large black-white score gaps carry negative views into evaluations of future cohorts of black students. 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. Negative first impressions lead teachers to be significantly less likely to over-rate but not more likely to under-rate black students' math and reading skills relative to their white classmates. Teachers' perceptions are sensitive to the lowest-performing black students in early classrooms, but non-responsive to highest-performing ones. This is consistent with the operation of confirmatory biases. Since teacher expectations can shape grading patterns and sorting into academic tracks as well as students' own beliefs and behaviors, these findings suggest that novice teacher initial experiences may contribute to the persistence of racial gaps in educational achievement and attainment.

JEL: I24, J15

*Rangel: Sanford School of Public Policy, Duke University; 262 Rubenstein Hall, Box 90312, Durham, NC 27708; Email: marcos.rangel@duke.edu

†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 bias in the US and international settings, less is known about the role of first impressions in shaping its extent and persistence. We study this in the education context by examining the early career steps of public school teachers. Our focus is on the extent to which first impressions feed into racial differences in teachers' expectations about students' capabilities. We build on 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 *nature* of this contact and by calling particular attention to the distribution of academic abilities of students within 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). The empirical literature on this topic lags behind, and our work aims at filling this gap.

Our analyses indicate that the distribution of academic abilities by racial group in a teacher's first classroom is a salient element in forming her beliefs and ultimately influences the way she evaluates students during the first three 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) containing subjective teacher assessments and blindly-scored standardized tests covering the same underlying skillsets. The objective test measure provides a reference point for diagnosing whether teachers are differentially lenient or strict for white vs. black students. Both white-black average score disparities and the presence of very low-performing black students in those early classrooms affect teacher evaluation patterns later on.

We find that teachers whose initial classroom contained black students performing substantially lower than white peers tend to exhibit reduced leniency for future cohorts of black students. An

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

increase of one standard deviation in the black-white performance gap among incoming students in a teacher's initial classroom corresponds to more than half the average racial gap in the evaluation of current students. Specifically, the worse the average performance of black students in a teacher's initial classroom (relative to white students), the less likely the teacher is to overrate the skills of her *current* black students (relative to current white peers). In contrast, the effects of early career experiences are more muted for teachers' relative stringency, or propensity to underrate black students in comparison to white peers.

The intensity of racial differences in teacher leniency is particularly sensitive to the racial composition of the *bottom tail* of her first classroom's ability distribution. In contrast, teachers' racial biases are not responsive to early exposure to high-performing black students. This asymmetry suggests that teachers are not updating their beliefs relative to a Bayesian uniform prior. Instead, their behavior is consistent with the presence of confirmatory bias, in which individuals assign more weight to information that confirms beliefs when faced with evidence (Rabin and Schrag, 1999). If teachers hold stereotypes about black students as low academic achievers, exposure to low-performing black students would affirm this prior more than is warranted by evidence, which in turn reinforces their biased rating of future black students.

In the process of establishing the impact of first impressions, we analyze in detail the extent of teacher racial bias in an elementary school setting. Our measure of bias is based on the careful comparison of teachers' evaluations of student reading and math ability with blindly-scored tests covering the same subjects. Our estimates yield significant white-black gaps even after accounting for differences on blindly-scored test performance, and teacher or classroom attributes. Teachers are more likely to exhibit leniency towards white students in both math and reading, while they are more likely to stringently assess black students in reading. Combining both subjects, we find that the average teacher is 2.3 p.p more likely to overrate white students relative to similarly performing black peers and 1.5 p.p. more likely to underrate black students. These results are not only statistically significant but also sizeable. As a reference, in this population, white students are overrated in 35% of cases, while they are underrated 18% of the time. Therefore, our base results

establish that there exist meaningful racial differences in teachers' subjective judgment of mastery within a given classroom.

Understanding teacher evaluation bias 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).[2] There is evidence that teacher expectations shape grading patterns and the propensity to steer students towards particular tracks, such as gifted and talented education (Donovan and Cross, 2002; Lavy, 2008; Burgess and Greaves, 2013; Botelho et al., 2015; Lindahl, 2016; Papageorge et al., 2016; Card and Giuliano, 2016). 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 indirect consequences of teacher racial biases. Evidence shows biases 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.[3] As a result, teacher evaluation biases may lead to gaps in attainment, school choice, future scholastic performance and, ultimately, occupational choices and labor market outcomes.[4] Efforts to bridge racial gaps in achievement and attainment can therefore benefit from a more informed understanding of this input.

[2]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.

[3]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.

[4]See Mechtenberg (2009) for a formalization of an argument like this. See also Lundberg and Startz (1983), who are explicit in modeling human capital investments' response to the presence of discrimination.

2 Related literature and contribution

Teachers are widely acknowledged to be an important 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.[5] 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.[6]

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 examinations.[7] A common approach is to juxtapose subjective teacher evaluations with blind assessments of student performance. Lavy (2008) capitalizes on the fact that students in Israeli high schools take two examinations covering same material and have the same format during their senior year, and that the grading of each exam happens under different anonymity regimes. Focusing on gender differentials, his findings indicate that male students receive lower marks in the non-blindly graded exams (relative to those blindly scored), and that these differences are larger than among girls. Blind/non-blind contrasts are also explored in a randomized control trial designed and implemented by Hanna and Linden (2012). The authors

[5] 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.

[6] 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.

[7] 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).

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.

Second, we examine the phenomenon of teacher overrating alongside underrating. A growing literature documents that differential outcomes by race can be driven by favoritism towards the majority group rather than by discrimination against the minority group (Greenwald and Pettigrew, 2014; Feld et al., 2016). Therefore, we examine whether teachers are less likely to overrate black students relative to their white peers, after accounting for blind-scored test performance and other individual characteristics.

Third and perhaps mostly 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. We hypothesize that the types and performance of students that teachers face in their initial classrooms may shape their expectations in subsequent years. This hypothesis is grounded in the recognition that beliefs about individuals and groups depend on the first piece of information to which one is exposed, and this first impression bias occurs in different contexts with lasting

consequences (Asch, 1946; Rabin and Schrag, 1999; Ambady and Skowronski, 2008). Alongside this work is research on stereotypes, theorized as representations of cross-group differences that allow more efficient mental processing of information (Hilton and von Hippel, 1996; Bordalo et al., 2016). In the education context, certain racial groups are associated with low academic achievement (Steele and Aronson, 1998; Alesina et al., 2018).[8] We trace the influence of teachers' early classroom experiences to examine how exposure to contexts that are consistent or inconsistent with stereotypes can influence teachers' subsequent assessments of black students relative to white peers. Rich data on course membership and the matching of teachers to students enable us to examine if characteristics of teachers' first classrooms – such as how average performance of students for a given ethnic group or the ethnicity of students at the extremes of the performance distribution – affect their assessments of future students belonging to the same racial group. Research exploring the origins of racial biases, in particular how they evolved in response to contexts that are more or less in keeping with negative prevailing stereotypes, are so far scarce. Our study examines these issues in K-12 education with a focus on the extent to which racial biases 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 the scope of elementary school teacher bias 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.

[8]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 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).

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. Personnel files also provide supplementary demographic information on teacher gender and race we utilize on our analyses.

Next we use course membership data to characterize multiple dimensions of new teachers' initial classroom experiences. In particular, we use achievement information from the first cohort of students (measured during the prior year before they interact with the teacher in question) faced by novice teachers to construct measures of each teacher's initial classroom conditions. As such, the initial ability of students in a fourth grade novice teacher's classroom is measured by their achievement in third grade standardized End-of-Grade (EOG) tests. This reliance on test scores which a given teacher could not influence precludes the inclusion of third grade teachers in our main analysis sample, since students are not taking standardized exams before that grade with the North Carolina system. We use these raw measures to compute group-specific summary statistics such as average scores and indicators for whether the highest or lowest scoring student in math or reading belongs to a given racial group in teachers' first classrooms. These variables, together with the shares of under-represented minority students by class, capture student composition and baseline ability distribution in each teacher's initial classroom.[9] The characteristics of novice teachers' first class comprise what we call "initial classroom characteristics." We match this information to the list of novice teachers and retain observations with non-missing initial classroom characteristics. We then track these novice teachers for up to three years in a fourth or fifth grade classrooms, after

[9]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 5179 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 that those in our sample were 3 percentage points less likely to be female and more likely to be white.

their initial year on the job.

To contrast these fourth and fifth grade teachers' perception of student abilities and students' actual performance, we rely on a section of NCERDC data available between 2007 and 2013. During these 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 for math and reading comprehension.[10] Levels 1 to 4 denote insufficient mastery, inconsistent mastery, consistent mastery, and superior performance, respectively.[11] 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.

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 at the local education agency level using software provided by the state Department of Public Instruction. The raw scores are then assigned to the same 1-4 achievement level scale described above summarizing the predetermined performance standards relative to grade-level comparisons.

Since EOGs are machine-scored using a common rubric, we consider these assessments of math and reading ability as "blind." Meanwhile, the fact that teachers know which student they are

[10]These assessments are used as an average at the state level to calibrate the Item Response Theory score distributions on EOG tests and are not used as inputs in teacher performance evaluations.

[11]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

evaluating and the race and ethnicity of each student renders their assessments "non-blind."[12] They provide assessments before knowing the students' actual test results. 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 told to focus on mastery over considerations of student behavior: *"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."*

As such the two measures of student skills are taken contemporaneously: standardized tests machine-scored and mapped to the four achievement levels of insufficient, inconsistent, consistent, or superior mastery and teachers' assessments of each student on the same scale. Since the measures anchor on the same underlying math and reading skills, we are able to use the correspondence between these two variables to derive evidence on the level of bias in teacher assessments.

The final dataset includes elementary students in a mixed-race grade 4-5 math or reading class who were taught by teachers with 1 to 3 years of experience (we classify novice teachers as having 0 years of experience) with nonmissing data on initial classroom conditions. In all, the analytic sample includes 2,677 teachers and their 125,520 unique students.

3.2 Descriptive statistics

Table 1 details teacher and students characteristics in the analytic sample. In contrast to near gender parity among students, nearly 9 out of every 10 elementary school teachers are female. The racial makeup is predominantly white, with 90% of teachers in this category while 8% are black. While the sample clearly skews towards white and female instructors, this is consistent with national

[12]Teachers have no incentive to be untruthful in their assessments. We understand that these subjective evaluations were introduced so that averages across the state could be used to calibrate item response theory methods to translate continuous blind-scores into the 4-level scale utilized by school authorities.

demographics of the teaching labor force.[13]

The NCERDC data includes student characteristics such as free and reduced lunch eligibility used to proxy for socioeconomic background and the number of annual absences we use to control for behavioral differences in the classroom. 30% of our sample of unique students are black, and 59% are eligible for free and reduced price lunch. Average math and reading scores in grades 4 and 5 are 0.12-0.15σ below the state average. The lower than average achievement records of students in part reflect the sample restriction to relatively inexperienced teachers.[14]

We then move to the central objective of the paper: to scrutinize the formation of racial biases, whose magnitude we hypothesize depend on teachers' initial classroom experiences. We are initially interested in exposure, or whether a teacher had at least one black student. Teachers who taught at least one black student during their first year may update their priors and identify performance signals in ways that are meaningfully different from those who had no previous contact with black students. Then, among teachers who had both black and white students in their initial classrooms, we examine the racial group-specific ability distribution. The extent to which the performance distributions of these two groups diverge from one another can shape subsequent expectations and biases.

We generate several measures to capture teachers' initial classroom experiences. Table 2 describes the features of early classrooms for all novice teachers in the sample. Math and reading classes share the same racial composition of 44% white and one-third black. Given the segregated nature of some schools and classes, it is possible that some teachers had homogeneous student populations during their first year that excluded African Americans. In Figure 1, over 11% of novice teachers belong in this category. Among those teachers with at least one black student during their initial year, there is substantial variation in student composition. There are classes ranging from

[13] 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.

[14] 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 not necessarily poses a threat to the internal validity of our estimates. We provide a more detailed argumentation in the next section.

just one black student to classes with only black students. Given that we need to juxtapose the ability distribution of two racial groups, for select analyses we retain only teachers with at least one black and white student in their first class.

The next variable in Table 2 is the 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. Black students lag behind white peers by 0.5-0.6σ in the typical classroom. Figure 2 (Panel a) plots the full distribution of white-black score disparities on the standardized math test. In the modal initial classroom, white students score over 0.5 standard deviations above black students, with the full set of observations approximating a normal distribution. Panel b of Figure 2 shows that the variation in white-black score disparities occurs across a range of early classroom compositions. It is not the case, for example, that the gap is increasing in the share of black students.

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 first create indicator variables for black students who are at the top and bottom of the classroom score distribution. This variable only turns on if black students exclusively occupy the top or bottom positions, not if both white and black students share the same relative position. According to Table 2, approximately one-fifth of early classrooms have a black student as their highest performer, while around one-half have a black student at the bottom of the class.

Expanding on this reasoning, we employ another set of variables to capture the non-overlapping tails of the ability distribution. We construct measures for the share of white (black) students whose scores are below the lowest-scoring black (white) student to describe the extent to which one tail under-performs the other. Conversely, we also examine the shares of white (black) students whose scores are above the highest-scoring black (white) student to describe the racial composition of

high-performers. Around one-quarter of black students in these initial classrooms scored lower than the lowest-performing white student, while 28% of white students tested above the highest-scoring black student in both subjects. In contrast, only 6% of black students scored above the highest white performer. Taken together, the evidence is consistent with a distribution of African American student test scores sitting to the left of the white student distribution across initial classrooms.

We then show some descriptive statistics suggesting the existence of racial disparities in teacher assessments 1-3 years *after* their first year as a novice teacher. Table 3 juxtaposes teacher assessments with blind-scored test performance on the achievement level scale of 1-4. For each standardized test achievement level, it displays the likelihood that the teacher assessment is higher, equivalent, or lower. To begin with, nearly half of students in both subjects demonstrate grade-level mastery at achievement level 3, with the bulk of remaining students either at levels 2 or 4. The distribution of student mastery in reading is lower relative to math.

Conditional on a level of performance, black students are consistently less likely to be over-rated and more likely to be under-rated relative to white students. The pattern holds true across math and reading. Among the highest performers at achievement level 4, for instance, teachers are 9 percentage points more likely to judge black students' mastery at a lower level in math and 14 percentage points more likely to do so in reading.

4 Empirical strategy

While the differences are illustrative, potential confounding factors imply that we cannot attribute gaps solely to teacher bias. For one, the distributions of student ability may vary across racial groups in ways not captured by coarse assessment categories. If white students have scores near the high end of achievement level 4 while black students cluster around the threshold between levels 3 and 4, for example, observed gaps in teacher assessments may reflect actual differences in underlying ability. Another factor is student-teacher assignment and the possibility that white

students may be disproportionately represented in classrooms with more lenient teachers. These are teachers who inflate assessments uniformly across all students. To ensure that student ability and teacher assessment practices are not explaining observed patterns, we turn to a regression framework which accounts for these sources of heterogeneity.

Our empirical approach defines the dependent variable as an indicator for teacher over-rating or under-rating. D_{irt} is the difference between non-blind ($NB \in \{1,2,3,4\}$) and blind ($B \in \{1,2,3,4\}$) assessments of student i, taught by teacher r in year t.[15] A teacher that consistently judges student mastery as higher than actual performance ($D_{irt} > 0$) is lenient, while the reverse signals stringency.

$$Pr(D_{irt} > 0) = \beta \text{Black}_{irt} + \alpha f(A_{irt}) + \mathbf{x}'_{irt}\Omega + \mathbf{z}'_{irt}\theta + \eta_{rt} + \epsilon_{irt}$$

The probability of lenient teacher assessments ($D_{irt} > 0$) for example, depends on several factors. The first is raw test scores A_{irt}, which enters flexibly into the model. $\mathbf{x}_{irt}$ encompasses covariates observed by both the teacher and econometrician (age, socio-economic status and gender), while $\mathbf{z}_{irt}$ covers covariates observed only by the teacher (and not by the econometrician). This distinction is relevant because, at least in principle, teacher assessments may take into account observations of student behavior that are correlated with race. Excluding $\mathbf{z}_{irt}$ therefore could lead to inconsistent estimates of the coefficient of interest β. We note two considerations that alleviate this concern. The first is the context of the teacher judgment measure, which *explicitly asks teachers to focus on evaluating student mastery over behavior*. Compliance then would imply that teachers are assessing students' math and reading competencies independent of behavioral attributes such as attendance and effort. To further address the concern over unobserved heterogeneity, we include days absent in the vector of student characteristics as a proxy for behavior (e.g. students who miss classes are the ones more likely to be rowdy when attending).

The inclusion of teacher-year fixed effects η_{rt} accounts for cross-teacher differences in assessment practices. Some teachers may be more strict or lenient in evaluations for all students, so this

[15] Each teacher is assigned one class per school-year.

estimation strategy removes these level effects. Since teacher-year corresponds to classroom-year indicators, our specification absorbs classroom-specific effects coming from the interaction of the teacher with a specific group of students, which may be factors that differentially impact the same teacher over different school years. Racial bias is then identified off within-classroom variation in assessments.

To determine the extent of racial differences in teacher assessments, we scrutinize the direction and magnitude of β in relation to the dependent variable. If the outcome is teacher overrating ($Pr(NB - B > 0)$), a negative coefficient on Black$_{irt}$ indicates that teachers are less likely to be lenient in evaluating black students relative to their white peers. One interpretation consistent with this finding is that teachers exhibit positive bias that favors white students. A positive coefficient for teacher underrating ($Pr(NB - B < 0)$ shows that black students are still more likely to be underrated compared to white students even after accounting for detailed student performance scores and behaviors. We take this differential stringency as evidence of teacher racial bias.

We then extend the model in order to scrutinize the formation of biases, whose magnitude we hypothesize depend on teachers' initial classroom experiences. We do so by augmenting our original empirical specification to estimate the effect of teachers' initial classroom conditions on subsequent biases. As before, the dependent variable is an indicator for leniency ($D_{irt} \equiv NB - B > 0$) or stringency ($NB - B < 0$), respectively.

$$Pr(D_{irt} > 0) = \delta \text{Black}_{irt} + \rho \text{Black}_{irt} \times (EC_r - \overline{EC}_r) + \gamma f(A_{irt}) + \mathbf{x}'_{irt}\Pi + \phi_{rt} + \epsilon_{irt}$$

We flexibly control for student i's mastery under teacher r in year t by including indicators for each raw score $f(A_{irt})$. As before, the vector of student covariates $\mathbf{x}_{irt}$ includes gender, socioeconomic status as represented by free and reduced lunch eligibility, and the number of days absent as a proxy for student behavior. We furthermore rely on teacher-year fixed effects to absorb classroom-level differences in attributes. This encompasses a substantive set of factors includ-

ing but not limited to teacher-level preferences for inflated or deflated ratings, classroom-specific shocks such as disruptions to learning, and year-grade level factors such as changes in testing regimes.

The model differs in its inclusion of an interaction term between Black_{irt} and measures of conditions in the teacher's first classroom (EC). We include multiple measures of those conditions summarized earlier, including whether the teacher was exposed to at least one black student during their first classroom, average white-black test score differences, whether the highest or lowest performing student was African American, and non-overlapping tails of white and black score distributions. We begin by estimating linear probability models using the pooled sample to replicate the magnitude of bias in both directions. Then we augment the specification with each early classroom measure to assess whether first impressions inform later teacher judgments. We eventually separate the sample into math and reading classes to better understand whether findings are driven by a particular subject. In addition to the main models, we also perform a number of robustness checks and examine the role of cross-subject early classroom attributes. For example, we test whether having a high-performing black student in math during teachers' novice year affects subsequent assessments of black students in reading.

The interpretation of the parameter on the interaction term can have an internally-valid causal interpretation under a particular assumption. The identifying variation in this case is that the early classrooms of novice teachers are not systematically assigned based on a predisposition for racial bias. This means that novice teachers are not selecting into schools and classrooms based on their own racial preferences or those of school administrators. We believe that this is too strong of an assumption, since the preferences of prospective teachers and employers likely matter for sorting across schools in the course of the job application/selection process. Yet we argue that this discretion is severely limited for both novice teachers and administrators for selection into classrooms *within a school*. We assume that administrators have no direct way of inferring racial bias predisposition among novice teachers hired by the school, and, just like teachers, have no access to information regarding classroom-level racial gaps in performance (once racial composition

is held constant) within a given school. We test this assumption by examining the relationship between teachers' observed characteristics and initial classroom characteristics with and without school fixed-effects. Table 4 presents evidence that this strategy can aid the identification of first impression effects. The odd columns employ variation within and across schools, while the even ones correspond to within-school variation only. Odd columns show that teacher demographic characteristics are strongly associated with initial classroom racial composition and performance by racial group. These relationships all become insignificant when conditioning on initial school fixed effects. Once we account for selection into schools, teacher characteristics simply do not hold a significant relationship with the characteristics of the initial classroom that we employ in our analyses.

Table 5 complements this finding by showing that the insignificant explanatory power of teacher characteristics is not due to lack of variation within schools. In fact, we see almost as much variation classroom-level racial gaps in previous years' test scores within schools as across them. While these do not guarantee the validity of the strategy, it provides substantial support to the idea that once allocated to a given school, novice teachers do not select into particular classrooms based on the dimensions of racial differences in student performance we care about. As a consequence, all subsequent models on the role of early classroom experiences use initial school fixed effects (interacted with student race). This is represented in our above specification by the demeaned measure of early-classroom (EC) relative to the early-school-level of the same conditions $(EC_r - \overline{EC}_r)$.[16]

Therefore, we do refer to the parameters we estimate as the *effects* of past racial gaps in performance on current evaluations of student mastery. While this design is internally valid, our findings cannot necessarily be extrapolated beyond the population of novice teachers. Due to data limitations discussed above, we have no direct way of assessing the impact of early experiences over more experienced teachers' evaluation patterns – or even to estimate the impact of learning over the reliance on priors informed by those early classroom assignments, which is one of the predictions of statistical discrimination models (Altonji and Pierret, 2001).

[16]This strategy is equivalent to running models including interaction of current-student race and fixed-effects for teacher's first assigned school.

5 Evidence on teacher bias

5.1 Extent of teacher bias

Table 6 reports findings on the likelihood of teacher overrating. We estimate a linear probability model for a pooled sample of grade 4 and 5 students in math and reading classes. In the base specification, a negative coefficient on Black_{irt} indicates within-classroom racial gaps in assessment even after adjusting for raw test scores and student behavior. Teachers are 2.3 percentage points less likely to overrate black students compared to white peers. To put this magnitude in context, teachers overrate 35% of the feasible set of white students, or those who record test scores up to achievement level 3.

The second and third specifications investigate whether teacher race and years of experience mediate the extent of bias. Previous studies establish certain advantages to having teachers who are demographically congruent with students, including higher expectations of student performance and attainment (Dee, 2005; Gershenson et al., 2016). As such we might expect black teachers to have an effect that attenuates the coefficient of interest. The second model includes an interaction term between black students and teachers, but the estimate is statistically indistinguishable from zero. The inflated standard errors indicate a relatively small sample of black teachers. As Table 1 shows, 8% of unique teachers in the sample are African-American. We similarly find no significant effects using an interaction term between black students and teachers with 2 or 3 years of experience, suggesting that classroom experience over this relatively condensed time period does not decrease the magnitude of bias in this sample. Note that we cannot rule out that teachers may learn to moderate their bias over a longer timespan due to the limited scope of our panel data.

We turn to Table 7 for the probability of teachers underrating student mastery. Relative to white students, teachers are 1.5 percentage points more likely to judge black students' mastery to be lower than their blind-scored test performance. This is a sizable difference, given that teachers underassess only 18% of white students. The nature of bias, then, extends beyond teachers' differential propensities for leniency by student race to include greater stringency for black students.

When exploring further with interaction terms on teacher race and experience, we again find that black teachers and those with greater tenure do not evaluate student mastery differently than their peers.

So far the results are based on pooled regressions involving students in both math and reading classes. Tables A1 and A2 separately report the extent of teacher bias by subject. We observe that teachers are 1.1 and 3.1 percentage points less likely to overassess black students in math and reading, respectively. The analogous estimates for underassessment are a statistically insignificant coefficient and 2.2 percentage points for math and reading. This underscores a form of asymmetry in the results: the extent of bias in both directions appears larger in reading than math. Results echo findings in related studies that find a larger quantitative effect in English relative to math (Lavy, 2008; Burgess and Greaves, 2013).[17] The difference may reflect the more subjective nature of reading or English language arts instruction, which leaves more room for interpretation relative to the problem-based nature of mathematics. Notably, we cannot reject that white and black students are similarly likely to be underrated in math. It is possible that racial disparities exist that are too small to be detected given the power of the present sample.

To establish bias, we assume that blind-scored standardized tests capture students' true underlying ability such that disparities between teacher assessments and raw scores arise from discriminatory behavior. However, there may be instances when the assumption is violated. These include stereotype threat in exam taking and cultural biases in exam design (Jencks, 1998). The former could bias our result towards finding a larger under-assessment of black relative to white students only if teacher-designed evaluations (exams, quizzes, problem sets) are more likely to affect minority students than the standardized tests administered by state-level education authorities. In this case, black students would consistently under-perform on teacher evaluations relative to their EOG scores – even with teachers who have an unbiased view of students. While we have no direct way of assessing this possibility, much of the literature on stereotype threat that takes place outside of

[17]Lavy (2008) found that Israeli teachers were most biased against males in English, although the corresponding effect in literature was smaller than math. Asymmetry in Burgess and Greaves (2013) only applies to the likelihood that teachers under-assess student mastery.

the lab setting involves high-stake exams, including those brought to school by an external agent and in an "one-shot" format, that are similar to EOG exams. Note that if teacher-designed evaluations are *less* likely to affect black students via stereotype threat than standardized tests such as EOGs, then our estimates are a lower bound on the magnitude of teacher racial bias. With respect to cultural bias the argument is similar. Biases could result from state-wide standardized tests that are more culturally neutral than classroom-level tests. Here, however, we argue that if these discrepancies emerge due to the discretion of teachers in formulating evaluations, we are comfortable in capturing them as an evaluation bias in the parameters of the model.

5.2 First impressions

Tables 8 shows the role of early classroom attributes on teachers' propensity to differentially exhibit leniency by race. The first column replicates racial differences in teacher overrating documented earlier for the pooled sample. Conditional on raw test scores, teachers are 2.3 percentage points less likely to overrate black students' mastery relative to white students. We test whether early exposure with black students mediates this effect. The coefficient on the interaction term between race and an indicator for having at least one black student in their initial class is positive and marginally significant. So we cannot reject that exposure attenuates racial disparities in assessment as proposed in a large literature within and outside economics.[18] Our point estimates indicate that being exclusively exposed to non-black students in the first classroom would more than double the bias in evaluation of current black vs. white students. Notice, however, that ultimately relatively few teachers initiate their careers in North Carolina without exposure to black students, which makes it hard to draw strong conclusions from these point estimates.

The specifications in columns 3 and 4 limit the sample to teachers who were exposed to at least one white and black student in their first class, and focus on the nature of that exposure. Relative to an early class in which black and white students had similar average scores, a one standard

[18] See Asch (1946); Lang (1986); Cornell and Welch (1996); Rabin and Schrag (1999); Ambady and Skowronski (2008); Pettigrew et al. (2011); and Devine et al. (2012).

deviation increase in relative score advantage among white students would decrease the likelihood of teachers overassessing later cohorts of black students by 1.3 percentage points (or more than 50% of the average bias). Thus initial impressions of white students' superior performance carry through to racial differences in future teacher assessment that cannot be explained by actual test performance.

Next we turn to an alternative expression of early classroom attributes: students at the tail ends of the performance distribution whose achievement may be particularly salient to novice teachers. We interact race with indicators for whether a black student earned the highest or lowest scores.[19] Having a black student as the lowest achiever in that first classroom decreases the likelihood of teachers overrating black students' mastery by 2.1 percentage points. Notably, teachers encountering highest-performing black students in their initial classrooms do not temper their bias later on towards this racial group. Instead, racial biases in assessments are accentuated when exposed to a low-performing African American student early on. These students may affirm any negative priors teachers hold about the group.

Table 9 expands on the relationship between the tails of early classroom ability distributions and how teachers subsequently assess students of different races. Column 1 replicates the last column from the previous table. Column 2 introduces a pair of variables on the share of black students below the lowest-scoring white student and the share of white students above the highest-scoring white student.[20] Moving from none to all black students scoring below the lowest-achieving white student exacerbates the racial gap in leniency by 2.5 percentage points. This suggests that teacher behavior is affected by the relative performance of the lower tail of the initial classroom distribution. This view is consistent with findings in Column 3, which shows that having more white students below the lowest-scoring black student in their first class significantly increases the likelihood that teachers will overassess future cohorts of black students.

We repeat these analyses in examining the phenomenon of underrating (Table 10). Lowest-

[19]This variable does not turn on in the case of ties between a black student and a student belonging to a different racial group.

[20]In combining these variables in a multiple regression setting we emulate a variation in the relative skewness of the black and white ability distributions.

performing and superstar black students in early classrooms do not appear to meaningfully influence future racial disparities in teacher stringency. Instead, a more salient factor is whether the teacher had black students in their first classroom assignment. Table 11 confirms that most outliers and tails of the performance distribution do not appear to shape future teacher stringency. While all coefficients are all in the expected direction, none of the four variables describing the extent of non-overlapping distributions are statistically significant.

Taken together, the evidence suggests some features of teachers' first classrooms do affect subsequent teacher evaluations, although the exact attribute and magnitude vary by context. Teacher overrating is sensitive to both the white-black average test score gap, and particularly to the presence of black low-achievers during the initial year. These two variables contribute significantly to any racial differences in overrating when teachers have between 1 and 3 years of experience. In contrast, these same factors play a less meaningful role in teachers' propensity to underrate student mastery. Another prevailing pattern is that early exposure to black academic superstars does not appear to influence teachers' subsequent assessment patterns, while the presence of black low-achievers does in select contexts.

In additional checks we provide evidence that these results are robust to a variety of specifications, including the exclusion of absences as a control variable, consideration of teacher attrition, measurement error in EOG scores, and the range of achievement levels included. Table A3 shows that excluding our proxy for student behavior, days absent, does not affect the magnitudes of observed coefficients on early classroom conditions. Next, we restrict the sample to only teachers who stayed for at least three years after their initial year, to determine whether our results are influenced by differential attrition. Table A4 shows first impression effects of similar magnitudes, as do the findings in Table A5 utilizing both lagged and contemporaneous EOG scores as controls for student ability. Table A6 acknowledges that because the achievement scale is bounded from above and below, at the lowest level only over-assessment can be observed, while the reverse is true at the higher end. Since blacks and whites have test scores unevenly distributed across these levels, it is possible that the incidence of over- and under-assessment we estimate here is an artifact of the

scale. However, Table A6 shows equivalent results to the ones presented above when we restrict the sample to students with measured achievement levels in which the under and over-assessment of performance are mathematically possible.

We furthermore examine the possibility that our results were driven by cases in which early classroom experiences were pretty close to total segregation (in which initial classrooms included very few black or very few white students). In Table A7 we re-examine the data considering only teachers with initial classrooms containing at least 20% of the students were either black or white. We find that in this case the qualitative conclusion is also not altered.[21]

Finally, we stratify the analysis by reading and math evaluation. Tables A8 to A11 show evidence that the prevalence of teacher leniency in both subjects is sensitive to the ability distributions of initial classrooms. For example, having a black low-performer early on decreases the likelihood of teachers overrating subsequent cohorts of black students by 2.1 and 3.3 percentage points in math and reading, respectively.[22]

6 Further discussion on belief formation and behavior

The asymmetrical influence of black low performers relative to high-achieving black students calls into question certain assumptions underlying a Bayesian updating model of belief formation. If individuals update their beliefs according to observed evidence, teachers should exhibit similar responsiveness to low-performing and high-performing black students.

The sensitivity to black representation near the lower end of the test score distribution calls to mind 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). Under this form of cognitive bias, individuals tend to misinterpret ambiguous evidence as

[21] Incidentally, one potential interpretation of the effects in this strata of early classrooms is that a more balanced racial composition, from the perspective of the teacher, improves the reliability of the measured racial gaps in performance.

[22] We also examine whether early impressions have cross-subject influence. That is to say, do the features of initial reading classrooms have spillover effects on teachers' future math assessment patterns and vice versa? Tables A12 and A13 suggest that effects tend to remain within the same subject.

confirming their priors or beliefs about the world.[23] 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."

In the context of this paper, novice teachers at risk of confirmatory bias can misinterpret evidence on the relative achievement of black vs. white students in their initial classrooms as supporting stereotypes about how students in a particular racial group should perform. Teachers' first impressions of lower average performance among black students or the presence of a particularly low-achieving black student can affirm the prevailing stereotype that black students are less academically inclined. This in turn can strengthen their belief in this stereotype more than warranted by evidence. In contrast, evidence countering this stereotype, such as having a stellar black student, is given less weight than would be expected under a rational agent. When confirmatory bias is sufficiently severe, theory predicts that learning through new evidence can exacerbate rather than bridge the existing bias (Rabin and Schrag, 1999).

Even though we present evidence that is consistent with confirmatory bias, we cannot fully rule out a model of learning based on a Bayesian framework in which teachers hold stereotypes against black students early on but update their beliefs as they learn about their students' true abilities. The reasons are twofold. First, 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. The second reason is that the classroom context differs crucially in one respect from statistical discrimination and learning studies in the workplace: interactions between a teacher-student pair usually last for only one year, specially 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

[23] 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).

context in which employers facing the same worker learn about productivity over a longer period and therefore rely less on race as a proxy for ability.[24]

7 Conclusion

We use statewide administrative data from North Carolina and document significant racial disparities in teacher assessment. Elementary teachers are less likely to judge black students at higher levels than their actual test performance compared to white peers who are observationally similar. These racial biases in leniency hold for both math and reading. There are analogous racial differences observed for teacher stringency, or underassessment. Black students are more likely to be assessed at lower levels of mastery in reading than white students in the same classroom after adjusting for test performance and student behavior. Underassessment in math competency is the only instance that lacks significant racial disparities in teacher judgment, although we cannot rule out that they do not exist among more experienced teachers or across different grade levels than those examined here.

We then carefully investigate the origins of these racial biases emerging early in a student's academic career. We hypothesize that novice teachers' early 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 disparities or the existence of superstar or lowest-performing black students in those initial classrooms influence subsequent assessment patterns. Evidence suggests that both attributes matter for teacher leniency later on. Having black students that lag on average white students' performance during a teacher's first year of experience reduces future levels of teacher leniency when assessing black students. Having a black student who scored the lowest in that initial class also makes a teacher less likely to overrate black students. In contrast, the effects of early career experiences appear more muted for teacher stringency during the first few years. Notably, novice teachers who are exposed to black superstar

[24]We examine a subset of our data restricting the sample to only teachers with the maximum level of experience are included. In Table A4 we see that our point estimates do not change in any meaningful way.

students early on do not appear to be affected. This asymmetry is compatible with the existence of confirmation bias, in which teachers holding stereotyped priors on low-achieving black students assign more weight to evidence that is consistent with stereotypes. These results are new to the literature, and call attention to the impact not only of exposure to racial groups but the nature of those interactions over the formation and reinforcement of racial bias in educational contexts.

We note several implications that follow from these findings. First, systemic biases against racial groups can adversely impact student performance, leading to under-investment in education for under-represented groups that in turn perpetuate longstanding achievement gaps. We argue that these effects can emerge either via reduced incentives to invest in education coming from cost-benefit perceptions or even via the tracking of students within education systems. As such, efforts to bridge gaps can benefit from a more informed understanding of this determinant of achievement disparities.[25].

Second, the finding that early classroom compositions and racial group-specific performance can shape future assessment practices implies a more deliberate approach to initial classroom assignment and professional development activities. This includes careful consideration of the consequences of assigning low-performing racial minorities to novice teachers. Teachers can also be made more aware of the ways in which their early interactions with such students can influence future expectations of minority groups. Above all, our findings point to the importance of bias training for teachers, with expected larger dynamic effects expected among novice members of the profession.[26]

Finally, a related point is that racial biases in assessments may be driven as much by teachers favoring a majority group as discriminating against racial minorities. We show that at least part of the overall racial disparity is due to teachers disproportionately overrating white students' abilities relative to their test-based mastery. Yet teachers may be less aware of the disparity-exacerbating effects of grading leniency if they are primarily focused on reducing differential stringency across

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

[26]Alesina et al. (2018) shows promising results of revealing results of implicit association bias tests to teachers as a way of combating negative stereotypes towards immigrant children in Italy, for example.

racial groups. Future work should further distinguish between these sources of bias and explore whether increased teacher awareness attenuates their effects.

References

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.

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

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.

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.

Feld, J., Salamanca, N., and Hamermesh, D. S. (2016). Endophilia or Exophobia: Beyond Discrimination. *The Economic Journal*, 126(594):1503–1527.

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.

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.

Greenwald, A. G. and Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist*, 69(7):669–684.

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.

Jencks, C. (1998). Racial bias in testing. In *The Black–White test score gap*, pages 55–85. Brookings Institution Press, Washington, DC, US.

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: An Official Journal of the Society for Personality and Social Psychology, Inc*, 9(2):131–155.

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.

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. Google-Books-ID: SdNOAAAAMAAJ.

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

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*.*

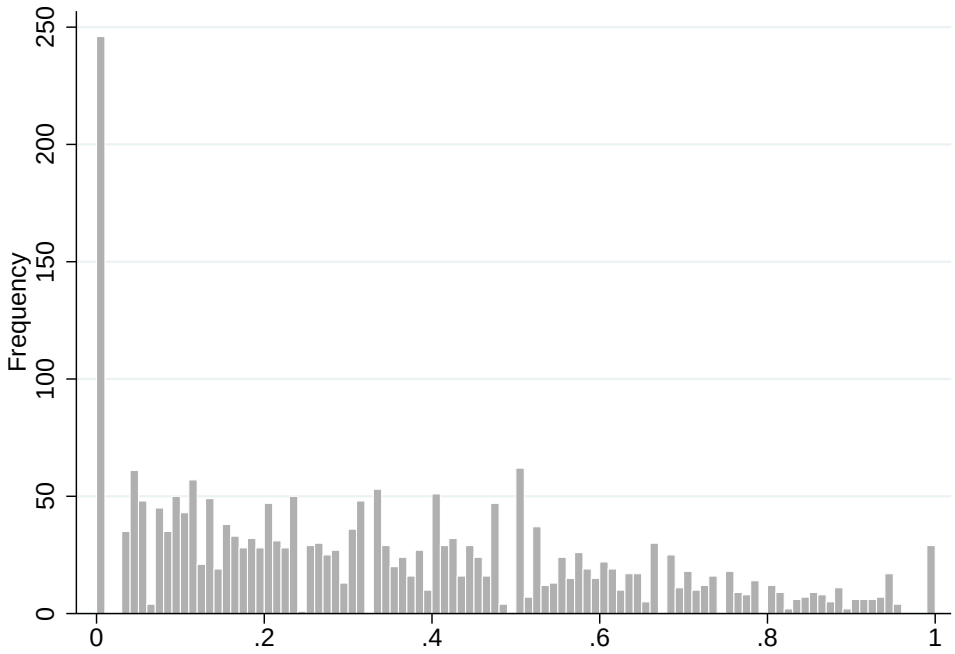
Sexton, P. C. (1961). *Education and income: inequalities of opportunity in our public schools*. Viking Press.

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.

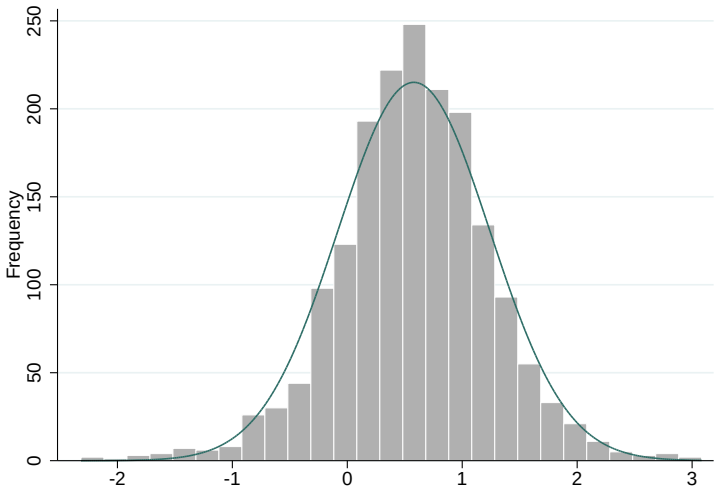
Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140.

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

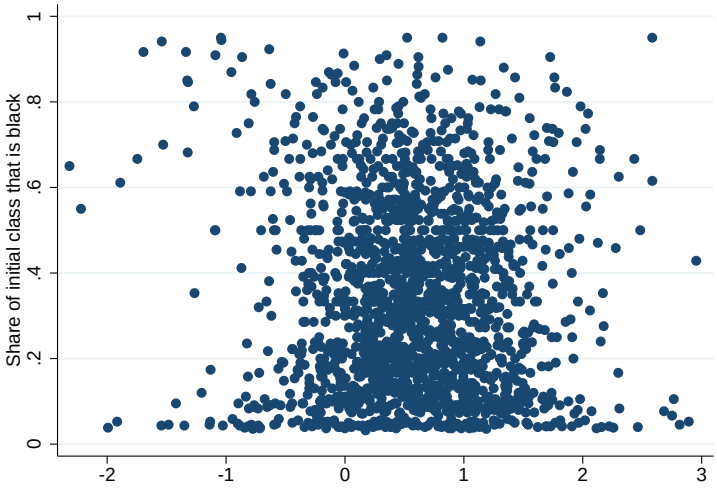


(a) Share of black students in early math class

Figure 1: Early classroom racial composition



(a) White-black gap in early classroom average math scores (σ)



(b) Early classroom composition vs. racial gap in achievement

Figure 2: Early classroom attributes

Table 1: Student and Teacher Characteristics

	Mean (1)	Standard deviation (2)
Student characteristics		
Female	0.49	0.50
White	0.48	0.50
Black	0.30	0.46
Hispanic	0.14	0.35
Asian	0.02	0.15
American Indian	0.02	0.14
Other	0.04	0.19
Econ disadvantaged	0.59	0.49
Grade 4 math scores	-0.14	0.98
Grade 4 reading scores	-0.15	0.99
Grade 5 math scores	-0.12	0.97
Grade 5 reading scores	-0.14	0.99
Grade 4 days absent	6.02	5.79
Grade 5 days absent	6.12	5.93
Observations	125520	
Teacher characteristics		
Female teacher	0.87	0.33
White teacher	0.90	0.31
Black teacher	0.08	0.27
Observations	2677	

Notes: Student sample includes unique individuals in grade 4 or 5 taught by someone who was a novice teacher (0 years of experience) from 2006-2012 and at present have between 1-3 years of experience. Test scores are normalized for all North Carolina students at year-grade level to mean 0 and standard deviation 1. Teacher sample includes unique individuals who taught math and reading with experience levels as described above.

Table 2: Initial Classroom Characteristics

	Math		Reading	
	Mean	Std. deviation	Mean	Std. deviation
	(1)	(2)	(3)	(4)
Share of initial class that is white	0.44	0.30	0.44	0.30
Share of initial class that is black	0.33	0.26	0.33	0.26
White-black gap in initial classroom test scores	0.58	0.66	0.53	0.69
Black student only has highest score	0.20	0.40	0.22	0.42
Black student only has lowest score	0.52	0.50	0.46	0.50
Share of white above highest-scoring black student	0.28	0.31	0.28	0.31
Share of white below lowest-scoring black student	0.08	0.19	0.09	0.19
Share of black above highest-scoring white student	0.06	0.16	0.06	0.17
Share of black below lowest-scoring white student	0.27	0.31	0.24	0.30
Observations	2300		2408	

Notes: This table shows characteristics corresponding to the initial classrooms of novice teachers from 2006-2012. We separate initial classrooms by subject. Observations shown correspond to the number of initial classrooms with non-missing data on racial composition.

Table 3: Teacher Assessment vs. Blind-Scored Achievement Levels

	Blind-Scored Achievement Level							
	Level 1		Level 2		Level 3		Level 4	
	White	Black	White	Black	White	Black	White	Black
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math								
Teacher assessment is higher	0.69	0.65***	0.45	0.38***	0.18	0.12***		
Teacher assessment is correct	0.31	0.35***	0.43	0.48***	0.64	0.63***	0.67	0.58***
Teacher assessment is lower			0.12	0.14***	0.18	0.25***	0.33	0.42***
Share of all observations	0.10		0.20		0.48		0.22	
Reading								
Teacher assessment is higher	0.73	0.69***	0.59	0.51***	0.28	0.19***		
Teacher assessment is correct	0.27	0.31***	0.35	0.40***	0.60	0.62***	0.69	0.55***
Teacher assessment is lower			0.06	0.08***	0.12	0.19***	0.31	0.45***
Share of all observations	0.16		0.24		0.44		0.16	

Notes: The teacher assessment variable is an indicator for teacher expectations of student mastery exceeding, equating, or below actual achievement. Sample includes black and white students in grade 4 or 5 taught by someone who was a novice teacher (0 years of experience) from 2006-2012. The sample restricts to students taught by teachers with between 1-3 years of experience. We exclude observations with missing student and teacher demographics, test scores, and teachers with missing early classroom attributes. Sample size is 65651 for math class and 78686 for reading class due to restricting to white and black students only. Statistically significant white-black differences are denoted using stars. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 4: Novice Teachers and Initial Classroom Characteristics

	Initial Classroom Characteristics					
	Any black student		Average score: black students		Black student has lowest score	
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Female teacher	-0.048**	-0.020	0.012	0.039	-0.093***	-0.050
	(0.019)	(0.017)	(0.036)	(0.046)	(0.032)	(0.052)
Black teacher	0.111***	0.012	-0.107**	-0.013	0.223***	0.014
	(0.023)	(0.007)	(0.043)	(0.054)	(0.038)	(0.059)
Non-white and non-black teacher	0.056	0.004	-0.018	0.011	0.111*	0.098
	(0.039)	(0.034)	(0.072)	(0.108)	(0.064)	(0.083)
School fixed effects		Y		Y		Y
Observations	2300	2300	2035	2035	2054	2054
Reading						
Female teacher	-0.047**	-0.015	-0.036	-0.010	-0.008	-0.000
	(0.020)	(0.016)	(0.038)	(0.045)	(0.033)	(0.048)
Black teacher	0.115***	0.006	-0.090**	-0.012	0.183***	0.004
	(0.024)	(0.009)	(0.043)	(0.047)	(0.038)	(0.054)
Non-white and non-black teacher	0.053	0.016	-0.095	-0.055	0.050	-0.061
	(0.043)	(0.036)	(0.081)	(0.141)	(0.070)	(0.113)
School fixed effects		Y		Y		Y
Observations	2408	2408	2107	2107	2131	2131

Notes: This table shows characteristics corresponding to the initial classrooms of novice teachers. We separate initial classrooms by subject. Even columns include initial school fixed effects, with standard errors clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table 5: Overall and Within-School Variation in Initial Classroom Characteristics

	Math	Reading
	(1)	(2)
Mean racial gap in average scores	0.58	0.53
Overall standard deviation	0.66	0.69
Within-school standard deviation	0.44	0.49
No. of teachers	1,784	1,863
No. of schools	763	770

Notes: The sample includes teacher-school-subject observations with non-missing racial gap in average scores for initial classrooms. We juxtapose the overall variation in the racial gap in average test scores in early classrooms with within-school variation. The within transformation subtracts the school mean from the racial-gap in scores and then computes the standard deviation.

Table 6: Teacher Overrating - Pooled Sample

	Teacher rating is higher: NB>B		
	(1)	(2)	(3)
Black	-0.023^{***}	-0.022^{***}	-0.020^{***}
	(0.003)	(0.003)	(0.004)
Black x Black teacher		-0.002	
		(0.009)	
Black x At least 2 years of exp			-0.005
			(0.005)
Observations	198,229	198,229	198,229
Teacher-school-grade-year FE	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices (0 years of experience) in 2006-2012 who currently have between 1-3 years of experience. The outcome is an indicator variable for teacher assessment exceeding actual achievement, and the table reports coefficients from linear probability models. We exclude observations with missing student and teacher demographics, test scores, and teachers with missing early classroom attributes. The omitted category in Column 3 is teachers with up to one year of experience. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, number of days absent, and teacher-school-grade-year fixed effects. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table 7: Teacher Underrating - Pooled Sample

	Teacher rating is lower: NB<B		
	(1)	(2)	(3)
Black	0.015***	0.015***	0.013***
	(0.003)	(0.003)	(0.003)
Black x Black teacher		−0.006	
		(0.007)	
Black x At least 2 years of exp			0.002
			(0.004)
Observations	198,229	198,229	198,229
Teacher-school-grade-year FE	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices (0 years of experience) in 2006-2012 who currently have between 1-3 years of experience. The outcome is an indicator variable for teacher assessment falling below actual achievement, and the table reports coefficients from linear probability models. We exclude observations with missing student and teacher demographics, test scores, and teachers with missing early classroom attributes. The omitted category in Column 3 is teachers with up to one year of experience. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, number of days absent, and teacher-school-grade-year fixed effects. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table 8: Overrating and Initial Classroom Characteristics - Pooled Sample

	Teacher rating is higher: NB>B			
	(1)	(2)	(3)	(4)
Black	−0.023***	−0.023***	−0.023***	−0.023***
	(0.003)	(0.003)	(0.003)	(0.003)
Black x Had black student in initial class		0.041*		
		(0.024)		
Black x Racial gap in average scores			−0.013**	
			(0.005)	
Black x Black student has highest score				0.006
				(0.008)
Black x Black student has lowest score				−0.021***
				(0.006)
Teacher-school-grade-year FE	Y	Y	Y	Y
At least one student of each race			Y	Y
Observations	198,229	198,229	157,995	157,995

Notes: The sample includes students taught by teachers who began as novices (0 years of experience) in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The outcome is an indicator for teacher assessment exceeding actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table 9: Overrating and Initial Classroom Characteristics - Pooled Sample (Cont'd)

	Teacher rating is higher: NB>B		
	(1)	(2)	(3)
Black	−0.023***	−0.023***	−0.022***
	(0.003)	(0.003)	(0.003)
Black x Black student has highest score	0.006		
	(0.008)		
Black x Black student has lowest score	−0.021***		
	(0.006)		
Black x Share of blacks below lowest-scoring white student		−0.025**	
		(0.012)	
Black x Share of whites above highest-scoring black student		0.003	
		(0.012)	
Black x Share of blacks above highest-scoring white student			0.025
			(0.020)
Black x Share of whites below lowest-scoring black student			0.038*
			(0.019)
Teacher-school-grade-year FE	Y	Y	Y
At least one student of each race	Y	Y	Y
Observations	157,995	157,995	157,995

Notes: The sample includes students taught by teachers who began as novices (0 years of experience) in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The outcome is an indicator for teacher assessment exceeding actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table 10: Underrating and Initial Classroom Characteristics - Pooled Sample

	Teacher rating is lower: NB<B			
	(1)	(2)	(3)	(4)
Black	0.015***	0.015***	0.015***	0.015***
	(0.003)	(0.003)	(0.003)	(0.003)
Black x Had black student in initial class		−0.054***		
		(0.021)		
Black x Racial gap in average scores			0.004	
			(0.005)	
Black x Black student has highest score				−0.002
				(0.007)
Black x Black student has lowest score				0.001
				(0.006)
Teacher-school-grade-year FE	Y	Y	Y	Y
At least one student of each race			Y	Y
Observations	198,229	198,229	157,995	157,995

Notes: The sample includes students taught by teachers who began as novices (0 years of experience) in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The outcome is an indicator for teacher assessment falling below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table 11: Underrating and Initial Classroom Characteristics - Pooled Sample (Cont'd)

	Teacher rating is lower: NB<B		
	(1)	(2)	(3)
Black	0.015***	0.015***	0.015***
	(0.003)	(0.003)	(0.003)
Black x Black student has highest score	−0.002		
	(0.007)		
Black x Black student has lowest score	0.001		
	(0.006)		
Black x Share of blacks below lowest-scoring white student		0.001	
		(0.011)	
Black x Share of whites above highest-scoring black student		0.007	
		(0.011)	
Black x Share of blacks above highest-scoring white student			−0.014
			(0.018)
Black x Share of whites below lowest-scoring black student			−0.003
			(0.018)
Teacher-school-grade-year FE	Y	Y	Y
At least one student of each race	Y	Y	Y
Observations	157,995	157,995	157,995

Notes: The sample includes students taught by teachers who began as novices (0 years of experience) in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The outcome is an indicator for teacher assessment falling below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Appendix

A Additional tables

Table A1: Teacher Overrating by Subject

	Teacher rating is higher: NB>B		
	(1)	(2)	(3)
Math			
Black	−0.011***	−0.011***	−0.012**
	(0.003)	(0.004)	(0.005)
Black x Black teacher		−0.001	
		(0.011)	
Black x At least 2 years of exp			0.001
			(0.006)
Observations	96,881	96,881	96,881
Reading			
Black	−0.031***	−0.030***	−0.025***
	(0.004)	(0.004)	(0.005)
Black x Black teacher		−0.004	
		(0.013)	
Black x At least 2 years of exp			−0.010
			(0.007)
Observations	101,268	101,268	101,268
Teacher-school-grade-year FE	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student and teacher demographics, test scores, and teachers with missing early classroom attributes. The outcome is an indicator variable for teacher assessment exceeding actual achievement. Omitted category is teachers with up to one year of experience in Column 3. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, number of days absent, and teacher-school-grade-year fixed effects. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A2: Teacher Underrating by Subject

	Teacher rating is lower: NB<B		
	(1)	(2)	(3)
Math			
Black	0.004	0.005	0.001
	(0.004)	(0.004)	(0.005)
Black x Black teacher		−0.010	
		(0.010)	
Black x At least 2 years of exp			0.005
			(0.006)
Observations	96,881	96,881	96,881
Reading			
Black	0.022***	0.022***	0.022***
	(0.003)	(0.003)	(0.004)
Black x Black teacher		−0.001	
		(0.008)	
Black x At least 2 years of exp			−0.000
			(0.005)
Observations	101,268	101,268	101,268
Teacher-school-grade-year FE	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student and teacher demographics, test scores, and teachers with missing early classroom attributes. The outcome is an indicator variable for teacher assessment falling below actual achievement. Omitted category is teachers with up to one year of experience in Column 3. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, number of days absent, and teacher-school-grade-year fixed effects. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A3: Robustness Check: Absences

	Teacher rating is:			
	Higher (NB>B)		Lower (NB<B)	
	(1)	(2)	(3)	(4)
Black	−0.016***	−0.016***	0.010***	0.010***
	(0.003)	(0.003)	(0.003)	(0.003)
Black x Racial gap in average scores	−0.014***		0.004	
	(0.005)		(0.005)	
Black x Black student has highest score		0.006		−0.002
		(0.008)		(0.007)
Black x Black student has lowest score		−0.021***		0.001
		(0.006)		(0.006)
Teacher-school-grade-year FE	Y	Y	Y	Y
Controls for days absent	N	N	N	N
Observations	157,997	157,997	157,997	157,997

Notes: This table tests for robustness to the omission of days absent as a control variable. The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. Outcomes are indicator variables for teacher assessments falling above or below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, and indicators for raw achievement scores. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A4: Robustness Check: Teacher Attrition

	Teacher rating is:			
	Higher (NB>B)		Lower (NB<B)	
	(1)	(2)	(3)	(4)
Black	−0.021***	−0.021***	0.016***	0.016***
	(0.004)	(0.004)	(0.004)	(0.004)
Black x Racial gap in average scores	−0.015**		0.003	
	(0.007)		(0.007)	
Black x Black student has highest score		0.003		0.005
		(0.010)		(0.009)
Black x Black student has lowest score		−0.025***		0.002
		(0.007)		(0.008)
Teacher-school-grade-year FE	Y	Y	Y	Y
Teachers with 3+ years of experience	Y	Y	Y	Y
Observations	104,585	104,585	104,585	104,585

Notes: This table examines whether the results are robust to restricting sample to teachers who stayed for at least 3 years. The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. Outcomes are indicator variables for teacher assessments falling above or below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A5: Robustness Check: Average of Lagged and Current Scores

	Teacher rating is:			
	Higher (NB>B)		Lower (NB<B)	
	(1)	(2)	(3)	(4)
Black	−0.008**	−0.008**	−0.000	−0.000
	(0.003)	(0.003)	(0.003)	(0.003)
Black x Racial gap in average scores	−0.009		0.002	
	(0.006)		(0.005)	
Black x Black student has highest score		0.002		−0.002
		(0.009)		(0.007)
Black x Black student has lowest score		−0.018***		−0.000
		(0.006)		(0.006)
Teacher-school-grade-year FE	Y	Y	Y	Y
Controls for average of lagged and current scores	Y	Y	Y	Y
Observations	149,202	149,202	149,202	149,202

Notes: This table tests for robustness to controls for test scores using both lagged and current scores. The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. Outcomes are indicator variables for teacher assessments falling above or below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, and indicators for the average raw achievement scores interacted with year and subject. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A6: Robustness Check: Achievement Levels

	Teacher rating is:			
	Higher (NB>B)		Lower (NB<B)	
	(1)	(2)	(3)	(4)
Black	−0.027***	−0.027***	0.017***	0.017***
	(0.004)	(0.004)	(0.003)	(0.003)
Black x Racial gap in average scores	−0.014**		0.006	
	(0.006)		(0.006)	
Black x Black student has highest score		0.010		−0.001
		(0.009)		(0.008)
Black x Black student has lowest score		−0.021***		0.002
		(0.007)		(0.007)
Teacher-school-grade-year FE	Y	Y	Y	Y
Achievement levels 2 and 3 only	Y	Y	Y	Y
Observations	130,492	130,492	136,014	136,014

Notes: This table examines whether the results are robust to restricting sample to students testing at achievement levels 1-3 for Columns 1-2 and achievement levels 2-4 for Columns 3-4. The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. Outcomes are indicator variables for teacher assessments falling above or below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A7: Robustness Check: Initial Classroom Composition

	Teacher rating is:			
	Higher (NB>B)		Lower (NB<B)	
	(1)	(2)	(3)	(4)
Black	−0.026***	−0.026***	0.022***	0.022***
	(0.004)	(0.004)	(0.004)	(0.004)
Black x Racial gap in average scores	−0.017**		−0.003	
	(0.008)		(0.009)	
Black x Black student has highest score		−0.002		0.001
		(0.010)		(0.010)
Black x Black student has lowest score		−0.021***		−0.002
		(0.008)		(0.007)
Teacher-school-grade-year FE	Y	Y	Y	Y
Early classes with 20%+ black and white students	Y	Y	Y	Y
Observations	78,637	78,637	78,637	78,637

Notes: This table examines whether the results are robust to restricting sample to early classrooms with at least 20% white and black students. The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. Outcomes are indicator variables for teacher assessments falling above or below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A8: Teacher Overrating and Initial Classroom Characteristics - Math

	Teacher rating is higher: NB>B					
	(1)	(2)	(3)	(4)	(5)	(6)
Black	−0.011***	−0.011***	−0.011***	−0.011***	−0.011***	−0.010***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Black x Had black students in initial class		0.025				
		(0.032)				
Black x Racial gap in average math scores			−0.030***			
			(0.008)			
Black x Black student has highest math				0.005		
				(0.011)		
Black x Black student has lowest math				−0.021**		
				(0.009)		
Black x Share of blacks below lowest-scoring white student					−0.065***	
					(0.019)	
Black x Share of whites above highest-scoring black student					−0.010	
					(0.018)	
Black x Share of blacks above highest-scoring white student						0.047*
						(0.028)
Black x Share of whites below lowest-scoring black student						0.056*
						(0.032)
Observations	96,879	96,879	77,366	77,366	77,366	77,366
Teacher-school-grade-year FE	Y	Y	Y	Y	Y	Y
At least one student of each race			Y	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The outcome is an indicator variable for teacher assessment exceeding actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A9: Teacher Overrating and Initial Classroom Characteristics - Reading

	Teacher rating is higher: NB>B					
	(1)	(2)	(3)	(4)	(5)	(6)
Black	−0.031***	−0.031***	−0.031***	−0.031***	−0.031***	−0.031***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Black x Had black students in initial class		0.051				
		(0.034)				
Black x Racial gap in average reading scores			−0.011			
			(0.008)			
Black x Black student has highest reading				0.002		
				(0.013)		
Black x Black student has lowest reading				−0.033***		
				(0.011)		
Black x Share of blacks below lowest-scoring white student					−0.009	
					(0.020)	
Black x Share of whites above highest-scoring black student					0.012	
					(0.019)	
Black x Share of blacks above highest-scoring white student						0.019
						(0.034)
Black x Share of whites below lowest-scoring black student						0.050
						(0.034)
Observations	101,268	101,268	80,607	80,651	80,607	80,607
Teacher-school-grade-year FE	Y	Y	Y	Y	Y	Y
At least one student of each race			Y	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The outcome is an indicator variable for teacher assessment exceeding actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A10: Teacher Underrating and Initial Classroom Characteristics - Math

	Teacher rating is lower: NB<B					
	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.004	0.004	0.003	0.003	0.003	0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Black x Had black students in initial class		−0.093***				
		(0.033)				
Black x Racial gap in average math scores			0.012			
			(0.008)			
Black x Black student has highest math				0.006		
				(0.012)		
Black x Black student has lowest math				−0.010		
				(0.009)		
Black x Share of blacks below lowest-scoring white student					0.010	
					(0.020)	
Black x Share of whites above highest-scoring black student					0.015	
					(0.018)	
Black x Share of blacks above highest-scoring white student						−0.012
						(0.029)
Black x Share of whites below lowest-scoring black student						0.020
						(0.034)
Observations	96,879	96,879	77,366	77,366	77,366	77,366
Teacher-school-grade-year FE	Y	Y	Y	Y	Y	Y
At least one student of each race			Y	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The outcome is an indicator variable for teacher assessment falling below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A11: Teacher Underrating and Initial Classroom Characteristics - Reading

	Teacher rating is lower: NB<B					
	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.022***	0.022***	0.024***	0.024***	0.024***	0.024***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Black x Had black students in initial class		−0.018				
		(0.021)				
Black x Racial gap in average reading scores			0.005			
			(0.006)			
Black x Black student has highest reading				−0.013		
				(0.009)		
Black x Black student has lowest reading				0.011		
				(0.008)		
Black x Share of blacks below lowest-scoring white student					−0.012	
					(0.015)	
Black x Share of whites above highest-scoring black student					−0.002	
					(0.016)	
Black x Share of blacks above highest-scoring white student						−0.039
						(0.025)
Black x Share of whites below lowest-scoring black student						−0.006
						(0.023)
Observations	101,268	101,268	80,607	80,607	80,607	80,607
Teacher-school-grade-year FE	Y	Y	Y	Y	Y	Y
At least one student of each race			Y	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. We exclude observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The outcome is an indicator variable for teacher assessment falling below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A12: Teacher Overrating and Cross-Subject Initial Classroom Characteristics

	Teacher rating is higher: NB>B			
	(1)	(2)	(3)	(4)
Math				
Black	−0.011***	−0.010***	−0.011***	−0.011***
	(0.004)	(0.004)	(0.004)	(0.004)
Black x Racial gap in average math scores	−0.030***	−0.019		
	(0.008)	(0.014)		
Black x Racial gap in average reading scores		−0.010		
		(0.011)		
Black x Black student has highest math			0.005	−0.011
			(0.011)	(0.019)
Black x Black student has lowest math			−0.021**	−0.024*
			(0.009)	(0.014)
Black x Black student has highest reading				0.011
				(0.014)
Black x Black student has lowest reading				0.004
				(0.010)
Observations	77,366	67,259	77,366	67,259
Reading				
Black	−0.031***	−0.029***	−0.031***	−0.028***
	(0.004)	(0.005)	(0.004)	(0.005)
Black x Racial gap in average reading scores	−0.011	0.004		
	(0.008)	(0.013)		
Black x Racial gap in average math scores		−0.009		
		(0.011)		
Black x Black student has highest reading			0.002	−0.011
			(0.013)	(0.020)
Black x Black student has lowest reading			−0.033***	−0.048***
			(0.011)	(0.016)
Black x Black student has highest math				0.012
				(0.015)
Black x Black student has lowest math				0.011
				(0.012)
Observations	80,607	66,048	80,607	66,048
Teacher-school-grade-year FE	Y	Y	Y	Y
At least one student of each race	Y	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. The sample also excludes observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The top panel uses an indicator variable for teacher assessment exceeding actual achievement, while the second panel uses an indicator for teacher assessment falling below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1

Table A13: Teacher Underrating and Cross-Subject Initial Classroom Characteristics

	Teacher rating is lower: NB<B			
	(1)	(2)	(3)	(4)
Math				
Black	0.003	0.002	0.003	0.003
	(0.004)	(0.005)	(0.004)	(0.005)
Black x Racial gap in average math scores	0.012	0.009		
	(0.008)	(0.013)		
Black x Racial gap in average reading scores		0.002		
		(0.010)		
Black x Black student has highest math			0.006	−0.013
			(0.012)	(0.019)
Black x Black student has lowest math			−0.010	−0.008
			(0.009)	(0.015)
Black x Black student has highest reading				0.016
				(0.013)
Black x Black student has lowest reading				−0.002
				(0.011)
Observations	77,366	67,259	77,366	67,259
Reading				
Black	0.024***	0.026***	0.024***	0.026***
	(0.003)	(0.004)	(0.003)	(0.004)
Black x Racial gap in average reading scores	0.005	−0.001		
	(0.006)	(0.010)		
Black x Racial gap in average math scores		0.006		
		(0.008)		
Black x Black student has highest reading			−0.013	0.001
			(0.009)	(0.013)
Black x Black student has lowest reading			0.011	0.026**
			(0.008)	(0.013)
Black x Black student has highest math				−0.015
				(0.010)
Black x Black student has lowest math				−0.011
				(0.009)
Observations	80,607	66,048	80,607	66,048
Teacher-school-grade-year FE	Y	Y	Y	Y
At least one student of each race	Y	Y	Y	Y

Notes: The sample includes students taught by teachers who began as novices in 2006-2012 who currently have between 1-3 years of experience. The sample also excludes observations with missing student demographics, absences, teacher demographics, test scores, and early classroom attributes. The top panel uses an indicator variable for teacher assessment exceeding actual achievement, while the second panel uses an indicator for teacher assessment falling below actual achievement. Initial classroom characteristics are demeaned using initial school-level averages. All specifications include student gender, socioeconomic status, indicators for raw achievement scores, and number of days absent. Standard errors are clustered at the teacher level. *** p<0.01, ** p<0.05, * p<0.1