Teacher Assessment Biases and Achievement Gaps in High School*

Alexis Orellana†

November 20, 2020

Abstract

In this paper, I study whether mis-assessment by teachers on subjective evaluations matters for students' educational outcomes. To do this, I employ administrative data from North Carolina's ninth-grade students between 2007 and 2013 that contains objective and subjective measures of achievement. Using these records, I first show that ninth-grade teachers are more likely to mis-assess girls and minorities. I then examine whether assignment to a teacher with a tendency to mis-assess, estimated from the longitudinal data, impacts end of high school outcomes. I find that exposure to teachers who are more likely to overrate students increases the intention to attend college, GPA scores, and SAT scores for minorities and girls, relative to other peers. These estimates are robust to the inclusion of teacher test score value-added. Since elementary teachers are also more likely to underrate minority students, these findings suggest an additional source of within-school disadvantage for this group of students, which persists at different educational levels.

JEL Codes: I21, J15, J16, J24

*I am indebted to Ronni Pavan for his guidance as well as to John Singleton for providing me access to the data housed at the North Carolina Education Research Data Center. I thank Lisa Kahn, Kegon Tan, Travis Baseler, and participants of the Applied Reading Group and the Student Seminar at the University of Rochester for helpful suggestions and feedback.

† Department of Economics, University of Rochester. E-mail: borellan@ur.rochester.edu

1 Introduction

For decades the concept of teacher quality has been primarily associated with the ability to raise test scores. Indeed, teacher quality's multidimensional impacts have only recently been the focus of much research (Jackson, 2018; Petek and Pope, 2018; Kraft, 2019). The scope of an effective teacher does not restrict to instruction, since teachers can also influence parental investment decisions (Kinsler and Pavan, 2020) and students' effort choices (Mechtenberg, 2009).[1] A necessary condition for parents to make informed decisions is that teachers provide accurate information about their students' performance. Yet, a growing number of papers show that teachers do not judge equally all students they serve. Girls and under-represented students are more likely to receive different ratings or grades relative to other observationally equivalent peers (Lavy, 2008; Hanna and Linden, 2012; Burgess and Greaves, 2013).

Research from psychology and economics has shown that discriminatory biases in evaluation can affect worker performance and productivity, and this mechanism can be particularly relevant in educational contexts.[2] Within schools, teachers' implicit biases or stereotyped perceptions could induce particular groups, e.g., girls or minorities, to perform worse relative to their peers in two different ways.[3] Firstly, teachers could convey their beliefs towards specific group performance through interaction with students (Keller, 2001). For example, a math teacher who thinks this subject is more difficult for girls could also challenge boys more frequently. Secondly, teachers could activate negative stereotypes held by students about themselves, leading to a self-fulfilling prophecy (Steele and Aronson, 1995; Spencer et al., 1999). In both situations, teacher beliefs about a specific group affect their performance.

Taken together, the available evidence in workplace contexts raises the question of whether some groups of students are more affected than others by exposure to mis-assessment and how relevant these effects are to explain the existing gaps in educational outcomes. In this paper, I investigate these two questions, exploiting the availability of objective and subjective measures of student achievement. Each student's contemporaneous test score defines an achievement level, which I consider as the objective measure. Simultaneously, teachers assess students using the same scale. Drawing on these two measures, I begin by analyzing whether ninth-grade teachers systematically mis-assess two observationally equivalent students of the same gender but different ethnicity. Then, I study whether teachers who are more likely to mis-assess students have a differential impact on girls and minorities, relative to other peers. To answer this last question, I employ a two-step ap-

[1] Recent papers studying the sensitivity of parental investments to better information are Cunha et al. (2013), Dizon-Ross (2019), Attanasio et al. (2019), and Bergman (2020).

[2] In sports, Price and Wolfers (2010) show that fouls are more likely to be called on players whose race differs from the NBA referee crew, while Parsons et al. (2011) find that strikes are more likely to be called when the umpire and pitcher are of the same race or ethnicity. Glover et al. (2017) show that biased managers negatively affect the performance of minority cashiers in French grocery stores.

[3] Throughout the paper, I consider a minority student to one who is either black or Hispanic.

proach. First, I define bias as the difference between the average score of the subset of peers rated by the teacher at the same level as the student and the student's test score. Based on this measure, I estimate each teacher's persistent propensity to overrate students, measured in test score standard deviations (s.d.). I refer to this quantity as teacher bias. Using a similar framework, I also estimate teacher test score value-added, which I employ as a proxy for teacher quality. Then, I project these teacher-specific estimates onto several outcomes measured in ninth and twelfth grades, such as contemporaneous test scores, intention to attend college, GPA, and SAT scores.

My estimation strategy requires the availability of teacher assessments and a comparable, objective measure of achievement for each student. In this regard, North Carolina's educational system is particularly well suited for these purposes. Between 2007 and 2013, the end-of-course exams included a question where teachers judged each student's achievement level. This assessment, reported at the moment students were taking the exam, is comparable with the achievement levels defined by the State Board of Education every year, based on the results of all students in the end-of-course tests. Simultaneously, it is also necessary to account for students' and teachers' non-random selection to schools and classrooms. For this reason, I estimate the extent to which each student is mis-assessed, controlling for individual ability and behavior lagged proxies, classroom-level characteristics, teacher fixed effects, and school fixed effects.

I find that North Carolina high-school teachers exhibit significant racial assessment biases. My preferred estimates show that, on average, teachers underrate minority students by 0.04 test score standard deviations relative to other-race students. On the contrary, they overrate girls on average by around 0.1 test score standard deviations. These estimates rely on within-classroom and within-teacher variation, controlling for any unobserved differences at the classroom level. Significantly, when a teacher and a student share the minority status, the average racial gap reduces to 0.02 test score standard deviations. English teachers are particularly inclined towards under-assessing minority students while math teachers over-assess girls by a higher amount.

Conditional on teacher test score value-added, I find that exposure to more biased teachers reduces contemporaneous test scores for all students but increases GPA, SAT taking, and SAT scores for minorities. At the same time, it also increases the intention to attend college reported by girls at the end of high school. Specifically, an increase of 1 s.d. in the teacher bias distribution reduces ninth-grade test scores on average by 0.009 s.d. and 0.004 s.d. for non-minority and minority students, respectively. This amount is equivalent to a decrease of roughly 0.2 and 0.1 standard deviations in teacher value-added. More biased teachers also influence students' expectations about college attendance in ninth grade. An increase of 1 s.d. in teacher bias leads to an increment of 0.5 percentage points in the likelihood of girls expecting to attend college. Regarding outcomes observed at the end of high school, my preferred estimates show that exposure to a more biased teacher increases GPA scores by 0.007 points, the likelihood of taking the SAT by 0.02 percentage points,

and SAT scores by 0.6 points for minority students. For girls, it also increases the intention to attend college declared for girls in twelfth grade by 0.04 percentage points. I compare these estimates with the impacts of increasing test score and behavior teacher value-added reported by previous work on the same outcomes. My estimates are comparable or higher in some cases to these impacts.

To address concerns related to selection of students into classrooms, I use two different strategies previously employed in the literature. First, I test whether my results are an artifact of selection based on observed variables. Then, I employ within-school, across-cohort variation to test selection based on unobserved variables. Additionally, I test whether the teacher fixed effects estimation is subject to measurement error. To address this issue, I also employ a split-sample instrumental variables strategy to correct for attenuation bias.

Regarding the relationship between teacher biases and their observed characteristics, I find that teachers improve their accuracy with experience. Nevertheless, they do so only by reducing the probability of overrating, while the probability of underrating does not associate with experience. I find that the probability of overrating is 1.8 percentage points (p.p.) lower for a teacher with five years of experience and 4.2 p.p. higher for a teacher with twenty years of experience. On the contrary, the estimates of experience on the probability of underrating are small and not statistically different from zero, suggesting that these biases are only partially mitigated through interaction with students. Moreover, I find that teachers who are more likely to underrate students have, on average, 0.1 s.d. lower in the value-added distribution, relative to more accurate teachers.

These findings are meaningful as they are informative about the long-term consequences of receiving negative ability signals in schools and how teachers contribute to the persistence or expansion of achievement gaps. Regarding the first point, since less-experienced teachers are more likely to serve disadvantaged students (Lankford et al., 2002), my estimates suggest that this group of students is also more likely to receive more biased signals from their teachers. They also highlight the importance of considering potential teacher biases when implementing programs to increase parental involvement, such as platforms that allow parents to view their children's grades.[4] Second, my results indicate that exposure to inaccurate teachers has heterogeneous effects for different groups of students. Advocates of teacher assessments as a feasible substitute for state-level or district-level mandatory tests argue that the former capture a broader range of abilities and are not determined solely by performance on a specific day. Nevertheless, these potential benefits do not consider that teachers are likely to base their judgments on gender and ethnicity. Finally, it is worth noting that my results understate this channel's total effect since I estimate the impacts of exposure on one specific grade. To the extent that these effects accumulate over time, persistent exposure to

[4]See Bergman (2020) for a recent field experiment which provided information to parents about their children's academic progress. Large school districts, such as Baltimore, Chicago, or Los Angeles, have adopted platforms, known as Learning Management Systems, to increase parent-teacher interactions.

under-assessment at multiple levels will expand their consequences.

The rest of the paper is organized as follows. In Section 2, I discuss how this paper situates in the current literature. Section 3 presents the institutional features of the North Carolina education system and the data. Section 4 introduces a simple framework incorporating teacher assessments into an education production function. I explain the empirical strategy in section 5. Section 6 and 7 contain the main results and robustness checks. Section 8 concludes.

2 Literature Review

In economics, the study of teacher assessment skills connects to two strands of literature.[5] The first one relates to understanding which skills drive teacher effectiveness. Scholars have focused on studying determinants of productivity such as subject-specific experience (Ost, 2014), specific job tasks (Taylor, 2018), peer-learning (Jackson and Bruegmann, 2009; Papay et al., 2020), or the quality of school-teacher matches (Jackson, 2013). While this set of papers focuses on test scores as the primary measure of productivity, there is evidence that teacher quality involves several other dimensions not captured by test scores (Jackson, 2018; Petek and Pope, 2018; Kraft, 2019). In this paper, I consider the skill to provide accurate assessments as one of these additional teacher quality dimensions.

The second body of research relates to the literature documenting the existence of implicit teacher biases, using information about blind and non-blind measures of children's academic skills (Lavy, 2008; Hanna and Linden, 2012; Cornwell et al., 2013; Burgess and Greaves, 2013; Botelho et al., 2015; Alan et al., 2018; Lavy and Sand, 2018; Terrier, 2020). These papers typically exploit the availability of a blind assessment (provided by an external grader or considering administrative test scores), and a non-blind teacher report to study gender or racial gaps in teacher assessments. A robust finding in this literature is that teachers overrate girls. In terms of race or ethnicity, significant differences also exist. For example, Hanna and Linden (2012) show evidence of systematic biases against low-caste students in India; Burgess and Greaves (2013) document that black Caribbean and black African students are underrated in England, relative to other-ethnicity peers; Botelho et al. (2015) find biases against black students in Brazil, and Alesina et al. (2018) show that math teachers with higher stereotypes give lower grades to immigrant students in Italy. In the U.S. context, Ouazad (2014) documents that elementary teachers overassess children of the same race. At the same time, Rangel and Shi (2020) find that elementary teachers are less likely to overassess black students, relative to observationally equivalent white peers.

[5] Other social sciences, especially educational psychology, have been interested in understanding how teacher knowledge of students relates to instruction and students' outcomes. See Hill and Chin (2018).

While the literature documenting racial and gender biases shows similar results in different countries and educational contexts, their consequences have been relatively understudied. Using a measure of implicit bias, the Gender-Science Implicit Association Test, Carlana (2019) shows that teachers with higher gender stereotypes increase the gender gap in math test scores and reduce the probability of girls attending more demanding high-school tracks. Another set of papers have exploited the availability of blind and non-blind assessments, as well as random assignment of teachers to classrooms, to study the effects of teacher biases on student outcomes (Lavy and Sand (2018) in Israel, Lavy and Megalokonomou (2019) in Greece, and Terrier (2020) in France). These three papers employ a measure of gender teacher bias at the classroom level, defined as the difference between boys' and girls' average gap between the non-blind and the blind score. While these papers show evidence of dissimilar impacts by gender, to the best of my knowledge, no prior studies link the existence of teacher assessment biases to racial achievement gaps, a particularly important topic in the U.S. educational system.[6] In this paper, I use a different approach to estimate teacher biases. Exploiting the panel structure of my sample, I estimate the persistent tendency to misrate students using a teacher fixed-effects regression.[7] This allows me to isolate a teacher-specific component from experience effects, student ability, and other potential determinants of teacher judgments, such as behavior. Moreover, while the papers above characterize teachers based on a measure of *relative* bias, gender favoritism, I estimate an absolute measure which allows me to test whether exposure to mis-assessment has dissimilar impacts across groups of students.

This paper also relates to the broader empirical literature studying the consequences of biases in human judgments. In the educational context, Diamond and Persson (2016) study the relation between test score manipulation and future labor market outcomes using data from Sweden, and Dee et al. (2019) analyze New York's high school exit exams.[8] This set of papers show that manipulation has impacts on several margins and, more importantly, these impacts vary by student and context. Crucially, in all of these cases, exams can be manipulated because these are not blind: teachers either see the student's identity, or the final score depends partially on subjective criteria, which give teachers the option to modify the results. This paper explores a different channel through which teachers' judgments can affect educational outcomes. I consider the difference between tests and assessments in ninth grade, where the impacts of teacher discretion on student outcomes are subtler than those generated by manipulating high-stakes tests at the end of high school. For instance, being underrated can change students' perceptions about their ability or change the probability of taking advanced classes.

[6] Papageorge et al. (2020) study how teachers expectations impact the probability of college enrollment. They exploit the availability of two teacher reports per student in the Education Longitudinal Study of 2002 to identify the causal effect of teacher expectations. Their results show that white teachers are less over-optimistic in their beliefs about black students' attainment.

[7] Among the papers mentioned above, only Lavy and Megalokonomou (2019) have data over multiple years.

[8] Recent empirical research in other areas about the importance of subjective assessments and their consequences are, for example, Hoffman et al. (2018) for hiring decisions, Li (2017) for grant evaluations, Frederiksen et al. (2020) for supervisors evaluations, and Parrado et al. (2020) for loan applications.

3 Institutional Background and Data

This section describes the institutional background of public schools in North Carolina and the features that make it a suitable context to study the relationship between teacher assessments and students' achievement in high school. My sample consists of all ninth-grade students in North Carolina's public schools between 2007 and 2013, obtained from the North Carolina Education Research Data Center. After describing the data, I present some descriptive patterns of racial and gender gaps in assessments.

3.1 North Carolina State Evaluation System

In the early 1990s, the North Carolina State Board of Education developed a School-Based Management and Accountability Program to improve student performance. Starting in the 1997-98 school year, North Carolina began testing high-school students by incorporating five end-of-course tests to the existing end-of-grade designed for students in third to eighth grades. Each year, students take a set of end-of-course tests to sample her knowledge of subject-related concepts according to the Standard Course of Study.[9] These exams are not graded by teachers, but scores count as 20% of a student's grade in the respective course.

3.2 Teacher Assessments

Between 2007 and 2013, each end-of-grade (third to eighth grades) and end-of-course (ninth to twelfth grades) test incorporated a question asking each teacher to assess students' achievement in the subject.[10] In particular, for all high-school students taking math and English courses. Table 1 describes the specific years in which these assessments are available for math and English. Teachers classified the achievement of their students in one of the following four categories:

- Level IV: Consistently performs in a superior manner and clearly beyond what is required to be proficient at grade-level work.
- Level III: Consistently demonstrates mastery of the grade-level subject matter and skills and is well-prepared for the next grade level.
- Level II: Demonstrates inconsistent mastery of knowledge and skills and is minimally prepared for the next grade level.

[9] In North Carolina, the subjects tested in high-school are English I, Algebra I, Algebra II, Biology, Civics, Chemistry, Geometry, Physics, U.S. History, and Political Science. Within this set, only English I, Algebra I, Algebra II, Biology, Political Science, and Geometry required teachers to assess their students for more than one year.

[10] These assessments were used to determine the cut scores that determine each achievement level, as well as an external variable to evaluate the validity of the tests (for a complete description of the standard-setting and the validity analysis, see North Carolina Reading Comprehension Tests Technical Report (2009), section 4.3, page 29, and section 7.3, page 61.)

- Level I: Does not have sufficient mastery of the knowledge and skills in the subject areas to be successful at the next grade level.

I consider the answer to this question as the teacher assessment of each student's achievement level. Figure 1 shows an example of the question included in the end-of-grade tests in 2011.[11] Two points are worth mentioning about its wording. First, it explicitly asks teachers to base their responses *solely on mastery* of the subject and provide information reflecting the achievement level uniquely. Second, each teacher has access to descriptors of the skills and aptitudes associated with each category. The precise wording of the question and the availability of information about the state-level achievement standards can alleviate concerns related to reference biases or whether other (unobserved) factors also influenced a judgment. Nevertheless, it could be possible that some teachers rate some students based on their behavior. I consider these potential concerns in the empirical analysis.

Objective Measures: The Department of Public Instruction sets standards of achievement for each student, using the same levels described above. Based on his end-of-course test score, each student is classified into one of these four levels. Table 2 shows the range of standardized scores for each achievement level between 2007 and 2012.[12] This objective level is available for each student with a valid test score, which I employ as a blind assessment. The availability of these two measures forms the basis to analyze the correlation of a biased assessment with future outcomes and whether classroom or students' characteristics influence teacher judgments.[13]

As mentioned before, these questions are available in the end-of-grade and end-of-course tests between 2007 and 2013. I choose to focus on ninth grade for two reasons. First, likely, the majority of students entering high school have not interacted with English or math teachers in previous courses, while in lower levels (such as elementary school), it is more presumable that teachers may have some level of information about these students from previous years.[14] This would be a concern if any student's unobserved characteristic has already influenced teachers' judgment in years not included in the data. Second, since most students take English I or Algebra I in ninth grade, I can match a high number of assessed students to their outcomes observed at the end of twelfth grade, which helps with the precision of the estimates.

I restrict the analysis to teachers with a valid certification and non-missing background variables in the School Activity Report (SAR) database. To match students and teachers, I employ a fuzzy

[11] Although Figure 1 corresponds to the text incorporated in the tests applied to students between third and eighth grades, the question used in the end-of-course tests is analogous.

[12] In 2013, the score intervals were the following: Level I included scores between 226-246; Level II between 247-252; Level III between 253-263, and Level IV between 264-281. There was no English I test during this year.

[13] See the description of academic levels for Algebra I in 2013 (page 63): https://files.nc.gov/dpi/documents/accountability/testing/technotes/mathtechreport1215.pdf

[14] Using the course membership data between 2006 and 2013, only 5% of all teachers who ever taught a ninth-grade class also did it in elementary or middle school grades.

matching algorithm, similar to the one used by Mansfield (2015) and Jackson (2018).[15] This procedure allows me to get high-quality matches and to avoid incorrect assignments if, for example, the person taking the test is not the teacher or if another source of coding imprecision exists. After this process, I match 85% of students for English I between 2007 and 2012; 70% of students for Algebra I between 2007 and 2013; and 90% for Geometry and Algebra II between 2007 and 2010.

3.3 Descriptive Analysis

I finalize this section by presenting some descriptive statistics about the final sample used in the estimation. I also describe the raw gender and racial assessment gaps observed in the data.

Summary Statistics: Table 3 presents descriptive statistics for the sample of students and teachers in the period 2007-2013. These data consider 459,253 student-year observations and 6,639 teachers in 507 schools. 51% of the students are male, and about 57% are white, 27% are black, 8% are Hispanic, and 2% are Asian. Regarding teachers, 54% are math teachers (Algebra I, Algebra II, or Geometry). The majority of them are white (85%) and female (75%), and the average years of experience is nine. 71% percent of them have a bachelor's degree. Finally, Table 3 shows that teachers rate most students as demonstrating a sufficient level of knowledge for the next grade level. Nevertheless, on average, only 51% of these assessments are aligned with those derived from the end-of-course test scores.

Descriptive Patterns: Figure 2 summarizes the main source of variation used in this paper. It plots the raw distribution of the difference between the teacher assessment (T_{ijst}) and the level associated with the test score (A_{ijst}), for Algebra I and English I between 2007 and 2013. As shown in Table 3, on average, teachers predict students' achievement correctly around 50% of times. While English teachers seem to be more accurate, the distribution is not symmetric, and, in both subjects, teachers tend to underrate students. While under-assessing students by one level of achievement ($T_{ijst} - A_{ijst} = -1$) occurs around 30% of times, over-assessing one level ($T_{ijst} - A_{ijst} = 1$) is less frequent, taking place between 15% and 20% of times.

To compute the unconditional gender and racial assessment gaps, I consider the following measure of bias. Based on the test score θ_{ijst} and the teacher assessment T_{ijst} observed for each student, I compute the difference between the average score of all students rated by teacher j in level T_{ijst} and student's i test score θ_{ijst}.[16] Then, I estimate regressions of the form:

[15] Specifically, I compute classroom-level background characteristics (total number of students, number of students by gender-race and grade cells) for each class observed in the end-of-course and the SAR databases. Then, I match classrooms based on a minimum distance algorithm. I refer the reader to the appendices in the papers above for details about how to implement the algorithm.

[16] I also employ this definition in section 5.1 to describe how I measure each teacher's persistent bias.

$$\overline{\theta}^{T}_{jst} - \theta_{ijst} = \alpha_1 + \sum_{v=2}^{20} \alpha^v \theta^v_{ijst} + \beta I_i + \sum_{v=2}^{20} \gamma^v \left(\theta^v_{ijst} \times I_i\right) + \epsilon_{ijst} \tag{3.1}$$

Where $\overline{\theta}^{T}_{jst}$ corresponds to the average score of all students rated by teacher j in the level T_{ijst}. θ^v_{ijst} is an indicator variable for whether student's i test score is in the v-th vigntile of the test score distribution, and I_i is an indicator variable for whether student i belongs to a specific group. I employ gender and ethnicity as the main categories to classify students. Figures 3 and 4 plot the coefficients $\beta + \gamma^v$ by race and gender, separately by subject. The upper plot in Figure 3 shows the average differences by gender, while the lower plot shows differences between minority (black-Hispanic) and white students. This plot shows that math teachers overrate girls across the entire distribution of test scores, and these differences are statistically significant at the 5% level. The average unadjusted gap corresponds to around 0.13 test score standard deviations. The lower plot shows that these differences are not statistically different from zero when we consider race. The opposite pattern occurs for English teachers. Figure 4 shows that, while there is evidence of gender and racial gaps in teacher assessments, the differences across race are more prevalent. The average gender and racial assessment gaps are 0.05 and -0.09 test score standard deviations, respectively. These patterns are consistent with previous literature documenting the existence of teacher biases. Since these gaps do not consider differences across school or background, I explore in Section 6 whether other teacher and student characteristics can explain these gaps.

One potential concern related to the nature of these assessments is that teachers may be providing uninformative reports. For example, if teachers do not take enough time to judge every student carefully, then T_{ijst} will not reflect what the teacher thinks about their students, leading to mis-classification. One way to test how informative these assessments are is to compute the joint distribution of teacher assessments and achievement levels. Tables 4 and 5 show that the correlation between the objective and subjective measures is high, supporting the assumption that teachers provide informative reports. Each column shows the distribution of teacher assessments conditional on the student's achievement level. For example, considering students whose test score corresponds to level III, 64% and 56% of all assessments are accurate for English and math, respectively. Finally, Figure 5 shows the distribution of valid assessments observed across all years in the sample for each teacher. Each plot's vertical red line represents the number of students rated by the average teacher in any subject. A math teacher rates around 48 students while an English teacher judges to 90 students on average.

4 Conceptual Framework

To motivate the empirical analysis, I consider a simple model incorporating teacher assessments into an education production function. In this extended setup, a teacher can improve skills through better instruction and induce effort by sending a signal to each student about their ability. To fix ideas, suppose that each student possesses an initial skill level θ_{i0} and an observable characteristic X_i. Between the initial and final periods, students acquire skills. I refer to the difference in skills $\Delta\theta_i = \theta_{i1} - \theta_{i0}$ as learning. I assume that learning generates by a combination of inputs and personal effort, according to the following specification, where e_i is the effort exerted by the student, ϕ_j^{VA} is teacher value-added, and τ_s are other school inputs:

$$\Delta\theta_i = \beta e_i + \alpha X_i + \phi_j^{VA} + \tau_s \tag{4.1}$$

Equation 4.1 considers that learning is an increasing function of effort. By simplicity, I assume that each student chooses an effort level based on his self-perceived skill level, which depends on the teacher's signal. The mapping $e_i = e_i(\theta)$ is linear and known to the student.

In addition to teaching, teachers provide assessments T_{ij} to their students. Assessing students is a costly task in terms of effort, and I consider two sources of heterogeneity for this cost. First, evaluate some groups of students correctly can be more demanding for some teachers. For example, based on previous experiences, teachers can consider that girls perform better than boys. These beliefs will make it harder for them to judge girls in later instances correctly. To incorporate this element, I assume that teacher beliefs about a student's skill depend partly on the student observed characteristics, $\theta_{i0} + \gamma X_i$. Second, teachers differ in their ability to assess students. I model this heterogeneity using a fixed parameter ϕ_j^B, which captures each teacher's bias to evaluate students. This parameter shifts the cost function so that teachers with $\phi_j^B < 0$ will optimally choose to underrate all students, regardless of their characteristics. Imposing $\phi_j^B = 0$ and $\gamma = 0$ implies that all teachers are unbiased.[17] Taking into consideration these points, I assume that each teacher chooses an assessment for each student, based on the following minimization problem:

$$\min_{T_{ij}} \quad \frac{(T_{ij} - (\theta_{i0} + \gamma X_i))^2}{2} - T_{ij}\phi_j^B \tag{4.2}$$

The optimality condition of this problem leads to the following assessment function:

$$T_{ij} = \theta_{i0} + \gamma X_i + \phi_j^B \tag{4.3}$$

[17] I model ϕ_j^{VA} as a common input that every student receives. Biased teachers could also impact students through heterogeneity in instruction, implying a correlation between ϕ_j^{VA} and ϕ_j^B. For example, they could interact differently with students or design evaluations reflecting her views about how difficult the material is for certain students (Keller, 2001). Unfortunately, I do not observe teacher practices in the classroom, so I abstain from incorporating this channel into the model.

Each student updates his belief about his skill level using the assessment T_{ij}. The updating process is a linear combination of the prior, which I assume is unbiased, and the teacher's signal. Students weight dissimilarly both signals, according to a factor $\pi_i = \pi(X_i) \in [0, 1]$:

$$\hat{\theta}_i = \pi_i \theta_{i0} + (1 - \pi_i) T_{ij} \tag{4.4}$$

After the student chooses the effort level $e(\hat{\theta}_i)$, the final skill level corresponds to $\theta_{i1} = \theta_{i0} + \Delta\theta_i$. Observed outcomes are a function of θ_{i1}:

$$y_i = \alpha^Y + \beta^Y \theta_{i1} + \epsilon_i \tag{4.5}$$

Substituting $e(\hat{\theta}_i)$ into (4.1) leads to a reduced-form equation which relates outcomes to the teacher characteristics, ϕ_j^{VA} and ϕ_j^B, as well as to the other inputs of the educational process:

$$y_{ij} = \delta_0 + \delta_1 \theta_{i0} + \delta_2 X_i + \delta_3(X_i)\phi_j^B + \delta_4 \phi_j^{VA} + \delta_5 \tau_s + \epsilon_{ij} \tag{4.6}$$

Where $\delta_1 = \beta^Y(1+\beta)$, $\delta_2 = \beta^Y(\beta\gamma(1-\pi_i)+\alpha)$, $\delta_3(X_i) = \beta^Y\beta(1-\pi(X_i))$, and $\delta_4 = \delta_5 = \beta^Y$. This simple model highlights how assessments can influence skill accumulation and later outcomes. From the teacher perspective, some students are more difficult to assess than others, and they also differ in their marginal cost of effort. As a consequence, students of the same ability who differ in observable characteristics receive different assessments. Then, depending on the weight students put to this signal, they change their self-perceived skill level and effort, impacting learning and outcomes.[18] The coefficient δ_3 in (4.6) comprises the total effect of exposure to a biased teacher. This coefficient depends on three parameters: (i) the return to skills (β^Y); (ii) the marginal productivity of effort (β); and (iii) the weight students put to the signal provided by the teacher (π_i). Since π_i is a function of the observable characteristic, the coefficient δ_3 varies with X_i, allowing the effect of being underrated or overrated to vary across different types of students. To the extent that the weight $\pi_i(X_i)$ is not constant, we expect to observe heterogeneous effects on students exposed to the same teacher. With this objective in mind, in the next section, I discuss my strategy to estimate teacher biases and their impacts across different groups of students.

[18] In related theoretical work, Mechtenberg (2009) employs a cheap talk game to study how teachers grading can influence gender differences in achievement and later outcomes. In her model, the grade sent by a teacher depends on the signal received by her and another teacher. Students update their effort cost based on this signal, but girls internalize it differently because they expect the teacher to behave differently depending on the student's gender. While her model is similar in spirit to my framework, in the sense that assessments may convey biased information used by students, there are some differences. First, I assume that teachers choose how to rate students based on their particular cost functions. Second, this model incorporates assessments into a standard education production function, including the effects of teacher quality. Third, this framework extends differences in gender to race and other observable characteristics summarized by X_i.

5 Empirical Analysis

My empirical strategy consists of two steps. In the first part, I construct estimates of each teacher's assessment practices, focusing on how frequently he predicts the academic level achieved in the test scores and how likely she underrates their students' achievement levels. The goal of this part is to isolate each teacher's permanent capability from the student's ability or behavior and other classroom-level characteristics that vary over time. The second step consists in projecting these estimates into different outcomes and test whether the estimated impacts vary by gender and race.[19]

5.1 Estimation of Teacher Biases

Let $T_{ijstc} \in \{1, 2, 3, 4\}$ be the teacher assessment of student i reported by teacher j in school s, year t, and subject c, and $A_{ijstc} \in \{1, 2, 3, 4\}$ the achievement level of the same student. Recall that A_{ijstc} is a deterministic function of the test score θ_{ijstc}, where each level is determined every year by the State Board of Education. T_{ijstc} and A_{ijstc} are discrete variables while θ_{ijstc} is a continuous measure, standardized to be mean zero and standard deviation one for each subject-year combination. Based on the test score and the teacher assessment, I construct the following measure of *bias*, defined as the difference between the average score of all students rated by teacher j in level T_{ijstc} and student i's test score:

$$\overline{\theta}^{T}_{jstc} - \theta_{ijstc}$$

For example, if an English teacher j rates student i in level 2, the bias corresponds to the difference between the average score of students in level $T_{ijstc} = 2$ in that year and course, and the actual score obtained by the student. Figure 6 displays the distribution of this variable separately by subject. At the end of these section I discuss other measures to characterize teacher biases using the discrete values T_{ijstc} and A_{ijstc}.

My approach consists of estimating each teacher's propensity to mis-assess students, accounting for non-random student selection into classrooms. With this objective in mind, I estimate the following regression, separately by subject:

$$\overline{\theta}^{T}_{jst} - \theta_{ijst} = X'_{ist}\gamma + C'_{ijst}\delta + \phi^{B}_{j} + f(\exp_{jt}) + \tau_s + \epsilon_{ijst} \tag{5.1}$$

In equation (5.1), X_{ist} considers student background characteristics (sex, race, parental education), a cubic polynomial of the seventh and eighth-grade test scores in math and language, days suspended out of school, and absences in seventh and eighth grades, and GPA in eighth grade. C_{ijst}

[19] Using the same data, Jackson (2014) and Bacher-Hicks et al. (2019) apply a similar two-step strategy to analyze the impacts of teacher quality and the impacts of school discipline policies on educational outcomes, although the latter constructs empirical Bayes estimates. I discuss the differences between these methodologies and my approach at the end of this section.

are leave-one-out, classroom-level, average characteristics of student i's peers (share of peers by race and gender; share of college-educated parents; average scores in math and reading in 8th grade; the average number of suspensions; share of repeating students). $f(\text{exp})$ is a flexible function capturing the effects of experience. I employ indicators for zero, one to two, three to five, six to ten, eleven to twenty, and more than twenty years of experience. The object of interest in this equation is ϕ_j^B. This fixed effect represents the persistent teacher bias after controlling for years of experience, student, and classroom characteristics. Finally, τ_s corresponds to a full set of school fixed effects.

The primary identification challenge in recovering estimates of teacher biases in equation (5.1) stems from non-random selection. Since students can sort into schools, and to teachers within schools, the comparison of mean differences at the teacher-level will not yield the differences in teachers' persistent biases. To address this source of bias, I assume that after controlling for a set of student-level and classroom-level variables, the allocation of teachers to students within a school is as good as random. Since ϕ_j^B is identified by comparing how different teachers assess observationally equivalent students in the same schools, the key identifying assumption I make is that, conditional on the school fixed effects and the controls X_{ist}, and C_{ijst}, unobserved characteristics of teachers and students are uncorrelated with assignment. Therefore, I make the following conditional independence assumption:

$$\mathbb{E}(\epsilon_{ijst}|\phi_j^B, X_{ict}, C_{ijst}, \tau_s) = \mathbb{E}(\epsilon_{ijst}|X_{ict}, C_{ijst}, \tau_s) \quad \forall j, \forall s$$

Under this assumption, conditional on the controls X_{ict}, C_{ijst}, the persistent bias of teacher j is uninformative about the expected characteristics of students taught by teacher j. Thus, the conditional difference in the outcome $\overline{\theta}_{jst}^T - \theta_{ijst}$ between teacher j and j' will yield the difference in the persistent bias between teacher j and j'. I present evidence to support the validity of this assumption in section 7.1. Previous studies using observational data have employed a similar assumption to uncover estimates of teacher quality[20]. I follow Jackson (2018) and incorporate lagged measures of test scores, suspensions, and attendance in seventh and eighth grades to account for potential selection in terms of ability and previous behavior. Accounting for these additional variables is particularly important in this context since teachers could also consider proxies of non-cognitive skills when assessing students. Furthermore, the classroom-level observed characteristics C_{ijst} account for sorting at the group level based on similar reasons.

A second empirical challenge consists of the presence of measurement error in the estimation of ϕ_j^B. Considering the discrete nature of the assessments T_{ijst}, any random shock that shifts this variable will generate mis-classification and thus lead to a biased estimate of the impact of teacher biases on student outcomes. I test whether my main results are robust to correcting for measurement error

[20]Previous studies typically consider two lags of test scores to account for selection based on ability (Rothstein, 2010; Jackson, 2014). Jackson (2018) also incorporates GPA and lagged measures of behavior to account for other sources of selection not captured by test scores.

in section 7.2.

Equation (5.1) allows to estimate each fixed effect using all the students observed in the sample for that teacher. Nevertheless, projecting the estimates $\hat{\phi}_j^B$ onto outcomes of the same group of students used to compute the fixed effects will lead to an endogeneity problem. If unobserved determinants of outcomes are correlated with unobserved determinants of teacher assessments and the estimates $\hat{\phi}_j^B$, this will lead to biased results. For this reason, I compute leave-year-out estimates of the persistent propensity to underrate students, $\hat{\phi}_{j,-t}^B$, which I employ as my measure of teacher behavior for students taught by teacher j in year t. Therefore, I analyze the association between a given student's outcome and her teacher's propensity to mis-rate based on a teacher measure which does not depend on the students observed in that year.[21]

One alternative to the specification (5.1) is the two-step empirical Bayes methodology, used extensively in this literature (Kane and Staiger, 2008; Chetty et al., 2014; Jackson, 2018; Petek and Pope, 2018; Bacher-Hicks et al., 2019). This method consists of calculating student-level residuals across multiple years and computing the average residual for each teacher. Then, each within-teacher average residual is shrunk using a factor that depends on the estimate of the true variance in teacher quality and the number of students. However, by construction, the computed residual is independent of the covariates, and any correlation between X_{it}, C_{ijst}, and the fixed effect ϕ_j will be captured by the former, potentially leading to understating the true variance of ϕ_j and biasing the final estimates. For these reasons, I estimate each teacher characteristic in one step employing the specification (5.1). The estimated fixed effect will capture the full impact of the teacher's persistent characteristic, rather than uniquely the component not predicted by the covariates X_{it}, C_{ijst}. Finally, I use a similar approach to recover estimates of teacher quality. I employ (5.1) to obtain estimates of teacher test score value-added, which I label ϕ_j^{VA}, using the end-of-course test scores as the dependent variable.[22] This set of estimates allows me to control for teacher quality in the second part of the estimation.

The estimated variance of the distribution of each teacher characteristic $\hat{\phi}_j$ is likely to contain sampling error due to small sample sizes. Consequently, the variance of the raw teacher fixed effects will probably overwhelm the signal and lead to overestimating the impacts on students' outcomes. To address the problem of sampling error, I follow Aaronson et al. (2007) and adjust

[21] I considered an alternative out-of-sample approach using time-invariant estimates from the period 2007-2013 and students from the period 2014-2015. This alternative is similar to the approach followed by Jackson (2014). While the main results are qualitatively similar, this strategy leads to several inconveniences. First, the number of students matched to teachers decreases because several new teachers enter schools every year. Second, the matching process is more inexact. After 2013, the end-of-course data does not identify the person taking the test, so the matching relies on information contained in the Course Membership data. On top of that, starting in 2014, the State Board of Education adopted a new standard course of study and a new accountability model, which implied changes in the names and codes of some math courses.

[22] This estimation strategy to estimate teacher quality has been previously used by Mansfield (2015).

this variance analytically. I assume that the estimated fixed effect decomposes into the true teacher effect ϕ_j, and an additive error ξ_j, where ξ_j is uncorrelated with ϕ_j:

$$\hat{\phi}_j = \phi_j + \xi_j \tag{5.2}$$

This assumption allows to isolate the variance of ϕ_j by subtracting the sampling variance $\text{Var}(\xi_j)$ to the variance of the fixed effects. Denote $\widehat{\text{SE}}(\hat{\phi}_j)$ to the standard error of the estimate ϕ_j obtained in 5.1. Then, the sampling variance is approximately the mean of the square of the standard errors:

$$\widehat{\text{Var}}(\xi_j) = \frac{1}{N}\sum_{i=1}^{N} \widehat{\text{SE}}(\hat{\phi}_j)^2 \tag{5.3}$$

Tables 6 and 7 present the raw and the adjusted variances of the distribution of each estimate $\hat{\phi}_j$, separately by subject. I find that around 10% of the raw variance is due to sampling error. Table 8 presents the correlation between the different estimates. The weak correlation between the two fixed effects suggests that teacher value-added does not associate with the capability of assessing students, and students exposed to low-quality students do not mechanically associate with more biased assessments. I find substantial variation in the estimates of $\hat{\phi}_j^B$ across teachers. Tables 6 and 7 show that the adjusted variance of $\hat{\phi}_j^B$ is 0.205 for math teachers and 0.109 for English teachers. These numbers imply that an increase of one standard deviation in the teacher bias fixed effect amounts to an over-assessment of 0.45 test score standard deviations for English teachers and of 0.33 test score standard deviations for math teachers. Figure 7 shows the distribution of the estimates of teacher bias ($\hat{\phi}_j^B$), and teacher value-added ($\hat{\phi}_j^{VA}$). Since a normal distribution can approximate the distribution of ϕ_j^B, an alternative interpretation of the variance of this distribution is that a student moved from the 50th to the 64th percentile will be over-assessed by around 0.4 test score standard deviations.

For ease of exposition in the out-of-sample estimation, I normalize each distribution of $\hat{\phi}_{j,-t}^B$ and $\hat{\phi}_{j,-t}^{VA}$ to have mean zero and standard deviation one, employing the adjusted variances shown in Tables 6 and 7. The next subsection discusses the second step of the empirical strategy.

5.2 Estimation of Impacts on Student Outcomes

After estimating and standardizing each teacher fixed effect, I estimate the following regression, which relates students outcomes with the leave-year-out estimates of teacher bias:

$$y_{ijst} = \alpha + \beta_0\hat{\phi}_{j,-t}^B + \beta_1\hat{\phi}_{j,-t}^B \times \text{Minority}_i + \beta_2\hat{\phi}_{j,-t}^B \times \text{Girl}_i + X'_{ijst}\gamma + C'_{ijst}\delta + \tau_{st} + \epsilon_{ijst} \tag{5.4}$$

In (5.4), the parameter of interest is $\beta = \{\beta_0, \beta_1, \beta_2\}$, which measures how outcomes change after increasing the teacher fixed effect $\hat{\phi}_{j,-t}^B$ by one standard deviation for different subgroups of stu-

dents. X_{ijst} and C_{ijst} correspond to the same controls employed in the initial step. To account for correlation in the outcomes of students assigned to the same teacher, I cluster standard errors at the teacher level. I consider several outcomes observed between ninth and twelfth grades. First, I employ the contemporaneous end-of-course tests of math and English I. Second, I use the intention to attend college declared by the student in ninth and twelfth grades. Finally, I employ information collected at the end of high-school regarding the GPA score computed by each school, SAT taking, and the SAT scores available for each student in the state records.

Although the fixed effect $\hat{\phi}_j^B$ is my primary measure of teacher bias, it is possible to characterize teachers' behavior using other alternative measures. In particular, I also recover teacher fixed effects after using each of the following three variables as the dependent variable in equation (5.1):

Under-assessment:

$$\mathbb{1}\{T_{ijst} < A_{ijst}\}$$

Over-assessment:

$$\mathbb{1}\{T_{ijst} > A_{ijst}\}$$

Precision:

$$|\overline{\theta}_{jst}^T - \theta_{ijst}|$$

The set of teacher fixed effects corresponding to each of the alternative measures of mis-assessment above has a different interpretation. *Under-assessment* and *over-assessment* characterize teachers according to their tendency to under-rate or over-rate students, regardless of how inaccurate the mis-assessment is. On the contrary, *precision* focuses on the magnitude of these differences. I discuss how the main findings relate to the use of each of these alternative measures in section 6.4.

6 Results

This section presents the main findings of the paper. In the first subsection, I analyze whether the racial gaps of Figures 3 and 4 are robust to the inclusion of teacher and student controls. Then, I examine whether teacher background characteristics and experience associate with the propensity to mis-assess students, measured as the teacher fixed effect ϕ_j^B in equation 5.1. Finally, I show the two-step estimation methodology results and analyze whether girls' and minorities' outcomes respond differently to more biased teachers.

6.1 Racial Gaps in Assessments

While Figures 3 and 4 show significant differences according to students' race and gender, the presence of confounding factors could drive these associations. For example, teacher practices,

student behavior, or sorting. To conduct this analysis, I employ the following specification, where I control for teacher experience and incorporate teacher and school-year fixed effects to account for time-invariant unobserved characteristics:

$$\bar{\theta}^{T}_{jst} - \theta_{ijst} = g(\theta_{ijst}) + X'_{ijst}\gamma + \phi_j + f(\exp_{jt}) + \tau_{ct} + \epsilon_{ijst} \tag{6.1}$$

In equation (6.1), the dependent variable corresponds to the difference between the average test score of students whom teacher j rated in the same level as student i, and the test score obtained by student i. Since this difference corresponds to test score standard deviations, the estimates ϕ^B_j will measure teacher bias using this scale.

$g(\theta_{ijst})$ is a third-order polynomial in the corresponding test score in ninth-grade. As part of the analysis, I consider several variables that could drive the differences in assessments, such as behavioral indicators or classroom composition. First, X_{ijst} includes student background characteristics (sex, race, and parental education), a cubic polynomial of the seventh and eighth-grade test scores in math and language, out-of-school suspensions and absences in seventh and eighth grades, and GPA in eighth grade. I control for unobserved classroom inputs by using classroom fixed effects (τ_{ct}). Therefore, this specification compares assessments of students exposed to the same teacher and the classroom inputs. In addition, $f(\exp_{jt})$ is a flexible function capturing the effects of experience. I employ indicators for different experience cells: 0, 1-2, 3-5, 6-10, 11-20, and more than 20 years.

Table 9 shows the estimates of biases for minority and girls, in test score standard deviations, relative to other peers. Column (1) shows that, after accounting only for contemporaneous test scores and unobserved teacher and classroom characteristics, minority students are rated on average 0.07 s.d. lower relative to other students. In contrast, girls are overrated by 0.09 standard deviations. The second row shows the estimate of an indicator equals to one when the student and teacher share the minority status. In this case, the mis-assessment reduces by 0.03 s.d. Column (2) includes the lagged ability and behavior controls for each student. In particular, the inclusion of a flexible polynomial on lagged test scores helps to account for potential mean reversion. After accounting for past test scores, the differences reduce to -0.04 s.d. for minorities, but it increases slightly to 0.1 s.d. for girls. Besides mean reversion, another potential concern relates to measurement error in the contemporaneous test score. It is possible that random differences in students' test scores close to the cutoffs that determine each achievement level A_{ijst} bias the estimates of the differences across subgroups. I use the same-subject test scores in eighth grade to instrument for the contemporaneous test score to account for this possibility. Specifically, I instrument the third-order degree polynomial $g(\theta_{ijst})$ using the polynomial of the eighth-grade test score. Column (3) shows that after accounting for measurement error, the coefficients attenuate slightly, but the differences remain statistically significant at the 1% level. I find that minorities are under-assessed by 0.02 test

score standard deviations. At the same time, girls are over-assessed by 0.09 test score s.d., relative to other peers in the same classroom. Finally, columns (4) and (5) reproduce the IV estimation separately by subject. The comparison of the estimates in each case shows that English teachers drive the bias for minorities. Conversely, math teachers exhibit a high tendency to overrate girls relative to boys.

An alternative way to quantify these patterns is to compute the differential probability of being underrated or overrated that girls and minority students face, relative to their peers. To conduct this analysis, I consider the measures of *under-assessment* and *over-assessment* defined at the end of section 5.2. Specifically, I run regressions of the form:

$$\mathbb{1}\{T_{ijst} < A_{ijst}\} = g(\theta_{ijst}) + X'_{ijst}\gamma + \phi_j + f(\exp_{jt}) + \tau_{ct} + \epsilon_{ijst} \tag{6.2}$$

Where the controls are the same as in (6.1). Table A1 in the Appendix presents the estimates for these two binary variables. Columns (1) and (2) replicate the results displayed in Table 9 employing my preferred measure of bias. Columns (3) and (4) present the estimates using under-assessment as the dependent variable. The least-squares estimates of column (3) show that minority students are 1.8 percentage points more likely to be under-assessed, while girls are 5.3 percentage points less likely to be under-assessed. After accounting for measurement error using the IV strategy, column (4) shows that these differences reduce to 1 p.p. and -4.5 p.p., respectively. A consistent pattern emerges if I consider the likelihood of overrating instead. Column (5) shows that minority students are less likely to be over-assessed by 1.5 percentage points. On the contrary, girls are 3.1 p.p. more likely to receive a higher rating, relative to peers of similar achievement and behavior in the same classroom. Accounting for measurement error changes these coefficients slightly.

Finally, I also consider another potential concern related to the discrete nature of the teacher assessments and the impossibility of observing biased assessments at the bottom and top levels of the test score distribution. If a teacher rates a given student in Level I, it is not possible to know whether this teacher is accurate or if she would have chosen a lower rating had it been available. A similar drawback follows in the case of Level IV. To address this concern, I restrict the sample to observations where it is possible to identify whether teachers mis-assessed the student's achievement. Table A2 shows the results of this estimation. Columns (1)-(2) employ observations in which the student's objective level belongs to Level II, Level III, or Level IV, and columns (3)-(4) restrict the sample to objective measures between Levels I and III. Overall, panels A and B show little difference with the main analysis of Table A1.

These estimates are comparable to those reported in previous literature studying teacher racial biases. Botelho et al. (2015) report that teachers underscore black students' grades by 0.02 standard deviations compared to white peers. In terms of binary outcomes, Burgess and Greaves (2013)

find that the probability of under-assessing increases by 2.5 and 3.5 percentage points for black Caribbean students in English and Science, respectively. Rangel and Shi (2020) find that elementary teachers in North Carolina are 1.5 p.p. more likely to underrate and 2.3 p.p. less likely to overrate black students. Interestingly, my estimates are very similar to those reported in their study. Taken together, they suggest that, at least in North Carolina schools, minority students undergo this source of disadvantage throughout elementary to high school courses. Moreover, since different teachers report these evaluations over time, they suggest a persistent teacher behavior pattern.

One potential explanation of gender assessment biases is that teachers have better perceptions of same-gender students. Dee (2007) finds that teachers' perceptions of students' performance increase when there exists a gender match. Since most teachers are female, this behavior could explain why girls tend to be favored. Students could also anticipate this behavior. An experiment run by Ouazad and Page (2013) shows that male students perceive that a female teacher will grade them less favorably than an external grader. While these studies have focused on gender differences, a similar explanation is feasible to explain racial disparities. To the extent that white teachers think that minority students will underperform relative to their white peers, they will be more likely to underrate them. Stereotyping beliefs could also drive this behavior. Rangel and Shi (2020) show evidence that racial differences in assessments reflect the existence of confirmatory biases, driven by the racial composition of the bottom-ability students during their initial years of experience. Although I do not have a direct measure of racial stereotyping beliefs[23], the estimate of the interaction between teacher and student minority status suggests that white teachers' beliefs could drive these associations.

6.2 Teacher Experience

Table 12 presents the estimates of the experience profile $\hat{f}(\text{exp})$ using the specification (6.1). I estimate the experience coefficients after pooling the sample across subjects, representing the association of teacher experience with the probability of observing a biased assessment for an average teacher. Each regression includes subject fixed effects in addition to the controls described in the previous sub-section. Using (6.1), I analyze how experience associates with the probability of under-assess and over-assess. The first row in Table 12 shows that, relative to novice teachers, higher experience does not decrease the probability of underrating students. At the same time, experience associates with the probability of overrating students. The second row shows that, relative to a novice teacher, a teacher with three to five years of experience will be three percentage points less likely to over-assess a student. A teacher with more than ten years will be five percentage points less likely to over-assess. Since teachers predict the objective achievement level around 50%

[23]For example, Carlana (2019) used the Implicit Association Test to measure teachers' gender favoritism, and Glover et al. (2017) employ the same test to measure managers' racial biases.

of times, these increases correspond to improvements of 6% and 10%, respectively. Teachers with more than twenty-one years of experience are slightly less likely to overrate. Thus, these estimates suggest that while experienced teachers provide more accurate reports than their novice counterparts, the probability of underrating does not decrease. Instead, the improvements derive from reductions in overrating. Tables A3 and A4 in the Appendix show these estimates splitting the sample by teachers' gender and race. The general patterns displayed in Table 12 are similar across these characteristics. In particular, no group displays statistically significant estimates for the association of experience with the probability of under-assessment. Also, minority teachers seem to decrease the probability of over-assessing students by a larger amount than their white counterparts.

6.3 Association Between Teacher Biases and Teacher Quality

Figure 7 plots the raw distribution of the fixed effect estimates $\hat{\phi}_j^B$. The adjusted variance of this distribution is 0.12, which implies, assuming a normal distribution, that a teacher located at the 16th percentile underrates students test scores by 0.35 s.d. while another teacher located in the 84th percentile overrates them by the same quantity. Using these estimates, I examine if there exists a relationship the teacher bias and value-added fixed effects. I analyze this association by running simple regressions of the form:

$$\hat{\phi}_j^B = \beta \hat{\phi}_j^{VA} + X_j'\gamma + \nu_j \tag{6.3}$$

In (6.3), X_j are teacher j's time-invariant characteristics (education, gender, and race), and $\hat{\phi}_j^{VA}$ corresponds to teacher value-added. Here, β corresponds to the average change in value-added associated with a 1 s.d. change in the teacher biases distribution. Since higher absolute values of $\hat{\phi}_j^B$ represent more biased teachers, I estimate these associations separately by first running (6.3) using values of $\hat{\phi}_j^B$ below zero and then using values above zero. Columns (1)-(3) in Table 11 show the association between teacher quality and teacher bias for teachers who are more likely to under-assess students, while columns (4)-(6) do the same for overrating teachers. Column (1) shows that, conditional on teachers below the 50th percentile in the teacher bias distribution, an increase of 1 s.d. in the test score value-added distribution associates with an increase of roughly 0.085 s.d. in the distribution of $\hat{\phi}_j^B$. This number implies that a teacher in the 16th percentile of the value-added distribution underrates students by 0.028 (0.085 x 0.33) standard deviations. Column (2) shows the association of $\hat{\phi}_j^B$ with teacher characteristics for teachers below the median of the distribution. This column shows that neither gender nor race is statistically different from zero. Only teachers with a degree higher than bachelor associate with a higher propensity to under-assess students at the 10% significance level. The coefficient of -0.228 suggests that teachers holding advanced degrees are more likely to be in lower percentiles of the teacher bias distribution, thus more likely to underrate students. Column (3) shows that the association between teacher quality and teacher bias in the lower part of the distribution is robust to including these controls. In this case,

the coefficient decreases to 0.07 and remains statistically significant at the 1% level. On the other hand, columns (4)-(6) show the estimates considering observations at the upper part of the teacher bias distribution. Column (4) shows that the association between teacher bias and teacher quality is not statistically different from zero for overrating teachers. Column (5) shows the association between teacher characteristics and the fixed effect in the upper part of the distribution. In this case, the female coefficient is negative and statistically significant at the 10% level, which suggests that female teachers are more likely to be in the central part of the distribution, which implies that the propensity to overrate is higher for male teachers. Finally, column (6) shows that the weak association between teacher quality and teacher biases does not vary after controlling for teacher time-invariant characteristics.

I also employ an alternative specification to examine how teacher quality varies with teacher bias across the entire teacher bias distribution. To do this, I regress teacher value-added onto indicators of quintiles of the distribution of $\hat{\phi}_j^B$, leaving the third quintile as the reference category and using the same observable characteristics as controls. I use this specification to examine whether this association holds for different sub-groups of teachers. Table A5 in the Appendix shows the estimates in this case. Columns (1) and (2) consider all teachers, while columns (3)-(4) and (5)-(6) show the estimates by teacher gender and race, respectively. Column (1) shows that the correlation between $\hat{\phi}_j^{VA}$ and $\hat{\phi}_j^B$ is particularly strong for those teachers with a higher propensity to under-assess students. The estimates imply that, relative to a teacher in the third quintile of the bias distribution, a teacher in the first quintile has 0.155 s.d. lower value-added. The remaining coefficients are statistically indistinguishable from zero, suggesting that the propensity to over-assess does not associate to teacher quality. Column (2) shows that this association holds after controlling for teacher gender, race, and education. I examine whether this pattern holds across different sub-group of teachers. Columns (3) and (4) split the sample by teacher gender, and the estimates suggest that women drive the negative correlation between value-added and the propensity to under-assess. Similarly, the comparison of columns (5) and (6) show that the group of non-white teachers also plays a role to understand the overall association.

To the best of my knowledge, only Lavy and Megalokonomou (2019) previously examined the relationship between teacher biases and teacher quality. Using data from Greece, they find that teachers who favor boys or girls in their ratings, relative to the opposite gender group, have lower teacher quality than more neutral teachers. Specifically, teachers who favor boys (girls) associate with a 0.04 (0.03) standard deviation reduction in the value-added distribution. My results indicate that teachers who have a higher propensity to under-assess students, regardless of gender or race, are negatively associated with teacher quality. Moreover, while this association is particularly strong in the lower-quintile of the teacher bias distribution, the association between teacher quality and teacher bias is not significant in the upper-quintiles. Considering the results of Table 11, my findings suggest that male teachers are more likely to overrate students and that education level

does not associate with lower biases.

6.4 Impact on Student Outcomes

Table 13 shows the results of the estimation of (5.4) for outcomes observed in ninth and twelfth grade. For ninth grade outcomes, I consider the score in Algebra I or English I end-of-course tests, and the intention to attend a two-year or four-year college after graduation. For twelfth grade outcomes, I employ the weighted GPA score reported by each school, intention to attend a two-year or four-year college, an indicator equals to one if the student took the SAT test after graduation, and the SAT score, conditional on taking the test. In each column, I report the estimates of the leave-year-out teacher bias fixed effect, ϕ_j^B, the interaction of this fixed effect with the female and minority status of the student, and the teacher value-added estimate ϕ_j^{VA}. Column (1) hows that, conditional on value-added, more biased teachers lead to a reduction in test scores. An increase of 1 s.d. in the distribution of ϕ_j^B associates to a decrease of 0.009 test score s.d. for non-minority boys (p-value<0.01). The interaction term between teacher bias and minority is 0.0056 (p-value<0.05) and implies a decrease of 0.004 test score s.d. for minority students. These coefficients are equivalent to roughly 20% and 10% of a decrease of 1 s.d. in teacher value-added. Column (2) shows the estimates for the intention to attend college declared at the moment of taking the end-of-course test. Conditional on teacher quality, the coefficients show that exposure to more biased teachers does not associate with minority students' expectations to attend college, but it does for girls. An increase of 1 s.d. in ϕ_j^B relates to an increase of 0.5 percentage points (p-value<0.05) in the likelihood of girls declaring an intention to attend college. This estimate is equivalent to an increase of around 1.7 s.d. in teacher value-added.

Columns (3)-(6) show the results for end of high school outcomes. Column (3) displays the association of teacher bias with the weighted GPA score reported by each school. In this case, more biased teachers are associated to relative increases in minority students' and girls' GPA. While the estimate of bias is not statistically different from zero, the interaction term is positive and statistically significant at the 1% level for minorities and at the 5% level for girls. Compared to the mean GPA score observed in the sample (3.18), this coefficient is small in relative terms (0.3% of the average value). Nevertheless, it is comparable to increases in teacher quality and very much in line with the expected subtlety of this type of bias. While an increase of 1 s.d. in the test score value-added distribution leads to an increase of 0.003 points (p-value<0.1) in GPA in twelfth grade, the interaction term of 0.0109 (p-value<0.01) implies that an increase of 1 s.d. in the teacher bias distribution for minority students leads to an increase of around 2 s.d. in teacher value-added.[24]

[24] Another way to interpret these numbers is to look at the effects produced by increases in behavior teacher value-added reported in the literature. Using a similar sample, Jackson (2018) finds that an increase of 1 s.d. in behavior teacher value-added increases GPA in twelfth grade by 0.021 points (p-value<0.01). His estimates for SAT taking and SAT scores are 0.012 (p-value<0.01) and -0.232 (p-value>0.1) (Table 7, page 2102). Using data from Los

Column (4) shows that the relationship found for girls' expectations in ninth grade persists three years later. Exposure to a teacher 1 s.d. more biased in ninth grade increases by 0.4 percentage points in the probability of declaring an intention to attend college in twelfth grade. For minorities, the interaction term is not statistically different from zero at the 10% level both in ninth and twelfth grades. Finally, columns (5) and (6) show the results for SAT taking and SAT scores. Similarly to the results in column (3), a higher value of ϕ_j^B induce relative increases in the probability of taking the SAT and SAT scores for minority students an girls. An increase of 1 s.d. in the distribution of ϕ_j^B leads to a 0.2 p.p. higher probability of taking the SAT (p-value<0.05) for minority students and girls. Conditional on taking the exam, exposure to more biased teachers associates with decreases for non-minority boys. An increase of 1 s.d. in teacher bias relates to a decrease of 1.42 points (p-value<0.01). Nevertheless, this decrease is lower for girls and the coefficient is positive for minorities. For the latter group of students the estimate corresponds to 0.6 points (p-value<0.1) in the overall SAT score.

My estimates are consistent with the findings reported in the existing literature about the gender-specific impacts of teachers who are more positively biased towards one group. Terrier (2020) shows that middle-school teachers who are *relatively* biased favoring girls increase their probability of selecting a scientific track in high school. Lavy and Sand (2018) find that primary teachers positively biased towards boys affect positively boys' test scores in middle and high school. Finally, Lavy and Megalokonomou (2019) also show that high school teachers who persistently favor boys also increase the probability of enrollment in a post-secondary program by 4 p.p. for boys and reduce this probability by 3 p.p. for girls. My estimates have a related, but different, interpretation. Conditional on being exposed to a teacher who is more likely to overrate *all students*, girls and minorities benefit relative to other peers. Since I employ an absolute measure of bias, one potential way to interpret the impacts of relative biases documented in the papers above is the following: a pro-boy teacher rates accurately boys but underrates girls. Since this teacher is, on average, more likely to underrate students, the differential effect experimented for girls from exposure to this teacher leads to worse relative outcomes for them. Analogously, if this teacher is pro-girl because she rates boys accurately but overrates girls, the heterogeneous impact of being exposed to a more over-assessing teacher will benefit girls relatively more.

I discuss how other measures of mis-assessment relate to student outcomes in Appendix A.1. I conduct a similar analysis characterizing teacher behavior using each of the variables defined at the end of section 5.2. Among this set of variables, the use of *precision* informs about how the magnitude of mis-assessment can also decrease girls' and minorities' outcomes. In this case, the teacher fixed effect captures how inaccurate the teacher is, regardless of whether she is, on average, more likely to underrate or overrate. Using the definition of bias, a teacher who underrates and

Angeles school districts, Petek and Pope (2018) estimate that an increase of 1 s.d. in behavior teacher value-added in elementary school raises the probability of taking the SAT by 1 percentage point (p-value<0.01) and GPA at the end of high-school by 0.013 points (p-value<0.01) (Table 5, page 51).

overrates by equal proportion will have a similar value of ϕ_j^B compared to another teacher who consistently rates students accurately. The use of precision helps to understand whether the size of the difference also matters. Table A10 shows that, on average, exposure to an imprecise teacher leads to lower outcomes for girls and minorities. The heterogeneity patterns are similar to those reported in Table 13, with the exception of contemporaneous test scores and SAT scores. For test scores, the teacher fixed effect is not statistically different from zero, while for SAT scores the interaction terms have a negative sign but they are not statistically significant at the 10% level.

7 Robustness Checks

7.1 Testing for Student Sorting

One natural concern is that principals can assign teachers based on unobserved characteristics to different classrooms. If this is true, then the estimates discussed in the previous analysis could merely reflect student sorting. One way to test this possibility is to project previous outcomes onto the estimated teacher fixed effects, conditional on the same set of controls as in (5.4), excluding skill and behavioral measures in 8th grade. This idea was implemented by Rothstein (2010) to test student sorting in elementary grades. I follow this approach using two sets of outcomes observed in 8th grade. Table 14 shows the estimates of the projection of the math and English test scores in 8th grade on each teacher fixed effect. Table 15 employs a similar specification using indicators for the number of days absent and whether the student received a suspension in eighth grade as right-hand side variables. The fixed effect estimates are small and not statistically different from zero for each table, supporting the selection-on-observables assumption.

Nevertheless, there is still the possibility of selection based on unobserved characteristics. For instance, students with particular characteristics not captured by ability or behavior proxies could be sorted towards the most inaccurate or lenient teachers. To test whether this type of selection could be driving the main results, I aggregate each teacher fixed effect at the school-year-subject level and rely on within-school, across-cohort variation in teacher assessment skills. The idea is to examine the impact of belonging to a cohort where, on average, teachers are more biased, compared to other cohorts with a different teacher composition in the same school. Since selection of students across cohorts based on these teacher characteristics is unlikely, aggregating the treatment at the school-year level helps to test the selection-on-unobservables assumption. This test has been used previously by Ost (2014) and Jackson (2014). If there is no sorting based on unobserved characteristics, the estimates obtained using this specification should be the same as the ones obtained using the specification 5.4. To do this analysis, I run the following regression, where $\bar{\phi}_{s,-t}$ represents the leave-year-out average value of $\hat{\phi}_{j,-t}$ in school s, and year t.

$$y_{ijst} = \beta_0 \bar{\phi}_{s,-t} + \beta_1 \bar{\phi}_{s,-t} \times \text{Minority}_i + \beta_2 \bar{\phi}_{s,-t} \times \text{Girl}_i + X'_{ijst}\gamma + C'_{ijst}\delta + \tau_s + \tau_t + \epsilon_{ijst} \quad (7.1)$$

Table 16 show the results of this test. The comparison of these estimates and the ones obtained using within-school-year variation indicates small differences. For example, the estimate of the interaction between the teacher fixed effect and the minority status of the student is 0.0109 (p-value<0.01) GPA using within-school-year variation, and 0.0101 (p-value<0.01) points using within-school, across-cohort variation. Similarly, while the estimates of the interaction for SAT taking and SAT scores are 0.0038 (p-value<0.05) and 2.0212 (p-value<0.01) using within-school-year variation, I obtain estimates of 0.0022 (p-value>0.1) and 2.2689 (p-value<0.01) exploiting within-school, across-cohort variation. A similar conclusion follows after comparing the estimate of the interaction of teacher bias and girls in the second row. Using within-school-year variation, the estimates for intention to attend college in ninth and twelfth grade are 0.0078 (p-value<0.01) and 0.0054 (p-value<0.01), respectively. After using within-school, across-cohort variation I obtain values of 0.0077 (p-value<0.01) and 0.0082 (p-value<0.01), respectively. The differences are also small after comparing the estimates for GPA, SAT taking, and SAT scores using both methodologies. Therefore, this analysis suggests that the main results can be attributed to exposure to more biased teachers and are not a consequence of selection of students to teachers, based on either observed or unobserved characteristics.

7.2 Measurement Error

The estimation of teacher bias can be prone to measurement error. To account for the possibility of classical measurement error in the estimates of $\hat{\phi}_j^B$, I employ an IV strategy. I split the sample into even and odd years and compute $\hat{\phi}_j^B$ for each sub-sample. Then, I use the same specification as in (5.4), instrumenting ϕ_j^B with the split-sample estimate. The assumption underlying the validity of this approach is that the estimation error is independent across the two sub-samples, but the estimates will be correlated since they capture the same teacher characteristic (Angrist and Krueger, 1995; Frederiksen et al., 2020).

Table 17 shows the estimates using this strategy. Overall, this table shows the same patterns observed in the main specification. After accounting for teacher quality, teachers who are more biased display a differential impact on student outcomes. While exposure to more biased teachers increase GPA, SAT taking, and SAT scores for minority students, they also increase the probability of girls expecting to attend college, both contemporaneously and at the end of high school.

8 Conclusion

In this paper, I study the relationship between teacher assessments received by ninth-grade students and future outcomes during high school using data from North Carolina public schools between 2007 and 2013. I employ information from the state-level standardized test scores and teacher reports about their students' achievement level to estimate the teachers' persistent tendency to provide biased assessments. This paper complements previous work documenting the existence of racial and gender teacher biases when evaluating students. It also connects to a broader literature studying the multidimensionality of teacher quality.

I document significant racial and gender differences in assessments for ninth-grade students in North Carolina. On average, teachers overrate girls by around 0.1 test score standard deviations and underrate minority students by 0.02 test score standard deviations. These differences derive from a differential probability of under-assessment that benefits girls but puts minority students at a disadvantage. Based on these patterns, I study whether teachers who persistently overrate students' achievement have heterogeneous impacts on their outcomes. I find that, conditional on a measure of test score value-added, teachers who are more likely to overrate have an additional positive effect on GPA scores, SAT taking, and SAT scores for girls and minority students. These estimates reflect the differential impact for this group of students conditional on being exposed to the same teacher. Also, I find that teachers who are more likely to overrate increase girls' reported intention to attend college. These estimates have comparable magnitude to the documented effects of increasing test scores and behavior teacher value-added.

This work leaves several avenues for future research. First, a natural question is related to how this teacher behavior impacts actual college attendance and major choice in the U.S. Recent work shows that teacher gender biases in high school predict performance in university admission exams and choice of fields of study (Lavy and Megalokonomou, 2019). It would be interesting to analyze whether biased assessments also affect post-school choices for minority students. Second, a critical unexplored channel in this paper is the role of parents. Unfortunately, the North Carolina data does not include explicit measures of parental beliefs or investments, which are meaningful drivers of human capital accumulation at this age, and also affected by the signals received from schools. Using experimental evidence, Bergman (2020) shows that improving the quality of school reporting or providing frequent information to parents about their child's effort in school induces gains in achievement. Therefore, improving teachers' ability to assess their students correctly is also of substantial importance when scaling this type of intervention.

References

D. Aaronson, L. Barrow, and W. Sander. "Teachers and Student Achievement in the Chicago Public High Schools". *Journal of Labor Economics*, 25(1):95–135, 2007.

S. Alan, S. Ertac, and I. Mumcu. "Gender Stereotypes in the Classroom and Effects on Achievement". *The Review of Economics and Statistics*, 100(5):876–890, 2018.

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

J. D. Angrist and A. B. Krueger. "Split-Sample Instrumental Variables Estimates of the Return to Schooling". *Journal of Business and Economic Statistics*, 13(2):225–235, 1995.

O. Attanasio, F. Cunha, and P. Jervis. "Subjective Parental Beliefs. Their Measurement and Role". Working Paper 26516, National Bureau of Economic Research, 2019.

A. Bacher-Hicks, S. B. Billings, and D. J. Deming. "The School to Prison Pipeline: Long-Run Impacts of School Suspensions on Adult Crime". Working Paper 26257, National Bureau of Economic Research, 2019.

P. Bergman. "Parent-Child Information Frictions and Human Capital Investment: Evidence from a Field Experiment". *Journal of Political Economy (Forthcoming)*, 2020.

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

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

M. Carlana. "Implicit Stereotypes: Evidence from Teachers' Gender Bias". *Quarterly Journal of Economics (Forthcoming)*, 2019.

R. Chetty, J. N. Friedman, and J. E. Rockoff. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood". *American Economic Review*, 104(9):2633–79, 2014.

C. Cornwell, D. Mustard, and J. Van Parys. "Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School". *Journal of Human Resources*, 48(1):236–264, 2013.

F. Cunha, I. Elo, and J. Culhane. "Eliciting Maternal Expectations about the Technology of Cognitive Skill Formation". Working Paper 19144, National Bureau of Economic Research, 2013.

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

T. S. Dee, W. Dobbie, B. A. Jacob, and J. Rockoff. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations". *American Economic Journal: Applied Economics*, 11(3):382–423, 2019.

R. Diamond and P. Persson. "The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests". Working Paper 22207, National Bureau of Economic Research, 2016.

R. Dizon-Ross. "Parents' Beliefs about Their Children's Academic Ability: Implications for Educational Investments". *American Economic Review*, 109(8):2728–65, 2019.

A. Frederiksen, L. B. Kahn, and F. Lange. "Supervisors and Performance Management Systems". *Journal of Political Economy*, 128(6):2123–2187, 2020.

D. Glover, A. Pallais, and W. Pariente. "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores". *The Quarterly Journal of Economics*, 132(3):1219–1260, 2017.

R. N. Hanna and L. L. Linden. "Discrimination in Grading". *American Economic Journal: Economic Policy*, 4(4):146–68, 2012.

H. C. Hill and M. Chin. "Connections Between Teachers' Knowledge of Students, Instruction, and Achievement Outcomes". *American Educational Research Journal*, 55(5):1076–1112, 2018.

M. Hoffman, L. Kahn, and D. Li. "Discretion in Hiring". *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

C. K. Jackson. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers". *The Review of Economics and Statistics*, 95(4):1096–1116, 2013.

C. K. Jackson. "Teacher Quality at the High School Level: The Importance of Accounting for Tracks". *Journal of Labor Economics*, 32(4):645–684, 2014.

C. K. Jackson. "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes". *Journal of Political Economy*, 126(5):2072–2107, 2018.

C. K. Jackson and E. Bruegmann. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers". *American Economic Journal: Applied Economics*, 1(4):85–108, 2009.

T. J. Kane and D. O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation". Working Paper 14607, National Bureau of Economic Research, 2008.

C. Keller. "Effect of Teachers' Stereotyping on Students' Stereotyping of Mathematics as a Male Domain". *The Journal of Social Psychology*, 141(2):165–173, 2001.

J. Kinsler and R. Pavan. "Local Distortions in Parental Beliefs over Child Skill". *Journal of Political Economy (Forthcoming)*, 2020.

M. A. Kraft. "Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies". *Journal of Human Resources*, 54(1):1–36, 2019.

H. Lankford, S. Loeb, and J. Wyckoff. "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis". *Educational Evaluation and Policy Analysis*, 24(1):37–62, 2002.

V. Lavy. "Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment". *Journal of Public Economics*, 92(10):2083 – 2105, 2008.

V. Lavy and R. Megalokonomou. "Persistency in Teachers' Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study". Working Paper 26021, National Bureau of Economic Research, 2019.

V. Lavy and E. Sand. "On The Origins of Gender Gaps in Human Capital: Short- and Long-Term Consequences of Teachers' Biases". *Journal of Public Economics*, 167(C):263–279, 2018.

D. Li. "Expertise versus Bias in Evaluation: Evidence from the NIH". *American Economic Journal: Applied Economics*, 9(2):60–92, 2017.

R. Mansfield. "Teacher Quality and Student Inequality". *Journal of Labor Economics*, 33(3): 751–788, 2015.

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

B. Ost. "How Do Teachers Improve? The Relative Importance of Specific and General Human Capital". *American Economic Journal: Applied Economics*, 6(2):127–51, 2014.

A. Ouazad. "Assessed by a Teacher Like Me: Race and Teacher Assessments". *Education Finance and Policy*, 9(3):334–372, 2014.

A. Ouazad and L. Page. "Students' perceptions of teacher biases: Experimental economics in schools". *Journal of Public Economics*, 105:116 – 130, 2013.

N. W. Papageorge, S. Gershenson, and K. M. Kang. "Teacher Expectations Matter". *The Review of Economics and Statistics*, 102(2):1–18, 2020.

J. P. Papay, E. S. Taylor, J. H. Tyler, and M. E. Laski. "Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data". *American Economic Journal: Economic Policy*, 12(1):359–88, 2020.

E. Parrado, A. Solís, R. Undurraga, and A. M. Montoya. "Bad Taste: Gender Discrimination in the Consumer Credit Market". Working Paper 10432, Inter-American Development Bank, 2020.

C. A. Parsons, J. Sulaeman, M. C. Yates, and D. S. Hamermesh. "Strike Three: Discrimination, Incentives, and Evaluation". *American Economic Review*, 101(4):1410–1435, 2011.

N. Petek and N. Pope. "The Multidimensional Impact of Teachers on Student Outcomes". Working paper, Department of Economics, University of Maryland, 2018.

J. Price and J. Wolfers. "Racial Discrimination Among NBA Referees". *The Quarterly Journal of Economics*, 125(4):1859–1887, 2010.

M. Rangel and Y. Shi. "First Impressions: The Case of Teacher Racial Bias". Working paper, 2020.

J. Rothstein. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement". *The Quarterly Journal of Economics*, 125(1):175–214, 2010.

S. Spencer, C. Steele, and D. Quinn. "Stereotype Threat and Women's Math Performance". *Journal of Experimental Social Psychology*, 35(1):4–28, 1999.

C. Steele and J. Aronson. "Stereotype threat and the intellectual test performance of African Americans". *Journal of Personality and Social Psychology*, 69(5):797–811, 1995.

E. S. Taylor. "Skills, Job Tasks, and Productivity in Teaching: Evidence from a Randomized Trial of Instruction Practices". *Journal of Labor Economics*, 36(3):711–742, 2018.

C. Terrier. "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement". *Economics of Education Review*, 77:101981, 2020.

9 Figures and Tables

Table 1: Availability of Teacher Assessments

	2007	2008	2009	2010	2011	2012	2013
Algebra I	✓	✓	✓			✓	✓
Algebra II	✓	✓	✓	✓			
English I	✓	✓	✓	✓	✓	✓	
Geometry	✓	✓	✓	✓			

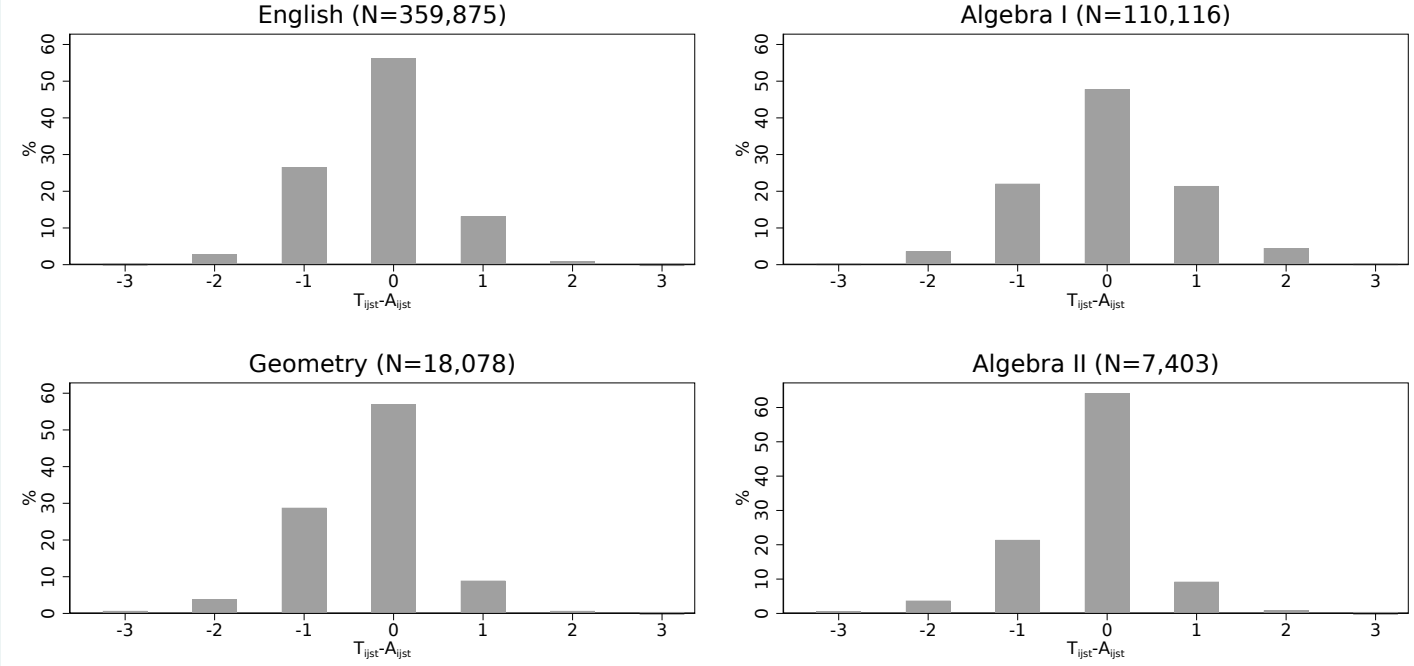
Table 2: Standards of Achievement, by Subject (2007-2013)

	Test Scores		
	2007-2012		2013
	Algebra I	English I	Algebra I
Level IV: Superior	158-181	157-176	264-281
Level III: Consistent	148-157	146-156	253-263
Level II: Inconsistent	140-147	138-145	247-252
Level I: Insufficient	118-139	119-137	226-246

Figure 1: Question Asking Teachers to Assess Each Student (End-of-Grade Test, 2011)

Information Requested	Column	Code (Fill In the Numbered Circle)
Achievement Levels for Mathematics *This coding requires input from the mathematics teacher who worked with the student during this school year. This is to be coded for <u>all</u> students who participate in end-of-grade mathematics.* *Instructions.* The mathematics teacher is to identify each student who, in the mathematics teacher's professional opinion, clearly and consistently exemplifies one of the achievement levels listed. If a student is not a clear example of one of the listed achievement levels, circle 9 in Column D is to be coded. The mathematics teacher should base this response for each student solely on mastery of mathematics. The mathematics 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 mathematics teacher's challenge is to provide information that reflects only the achievement of each student in the subject matter tested. The mathematics teacher should therefore rely chiefly on professional experience about what is the appropriate achievement level.	**D**	1 = Achievement Level I 2 = Achievement Level II 3 = Achievement Level III 4 = Achievement Level IV 9 = Not a clear example of any of these achievement levels *See Appendices A1–A6 in this manual for descriptions of the four mathematics achievement levels at grades 3–8.*

Figure 2: Differences Between Assessments and Achievement Across Subjects



Notes: Each plot shows the frequency of the difference between the teacher assessment T_{ijst} and the achievement level A_{ijst} in the respective subject, based on the total number of assessments available in the sample between 2007 and 2013.

Table 3: Summary Statistics.

Variable	Mean	Std. Dev.	Observations
Unit of observation: Student-year			
White	0.55	0.50	462815
Black	0.25	0.43	462815
Hispanic	0.08	0.27	462815
Asian	0.02	0.14	462815
Algebra I Score	0.24	0.95	252647
English I Score	0.21	0.91	388063
Reading score (8th grade)	0.08	0.95	462815
Math score (8th grade)	0.09	0.95	462815
Repeated (8th grade)	0.01	0.08	459561
Suspended out of school (8th grade)	0.28	1.70	396374
Repeated (7th grade)	0.01	0.09	462815
Suspended out of school (7th grade)	0.16	1.40	462815
Days Absent (7th grade)	6.68	6.78	462815
Times Tardy (7th grade)	0.95	4.06	462815
Teacher judgment (Math): Level I	0.10	0.30	145133
Teacher judgment (Math): Level II	0.22	0.42	145133
Teacher judgment (Math): Level III	0.45	0.50	145133
Teacher judgment (Math): Level IV	0.23	0.43	145133
$\mathbb{1}(T_{ijts} = A_{ijts})$ (Math)	0.50	0.50	145133
Teacher judgment (English): Level I	0.05	0.21	367070
Teacher judgment (English): Level II	0.18	0.38	367070
Teacher judgment (English): Level III	0.53	0.50	367070
Teacher judgment (English): Level IV	0.24	0.43	367070
$\mathbb{1}(T_{ijts} = A_{ijts})$ (English)	0.56	0.50	367070
Unit of observation: Teacher			
White teacher	0.84	0.37	6366
Black teacher	0.14	0.35	6366
Hispanic teacher	0.01	0.08	6366
Female teacher	0.75	0.43	6366
Experience	9.74	9.87	6360
Initial experience	8.65	9.65	6358
Education: Bachelor's degree	0.72	0.45	6360
Education: Master's degree	0.27	0.44	6360

Table 4: Joint Distribution of A_{ijst} and T_{ijst}: English

	Achievement Level (A_{ijst})			
	Level I	Level II	Level III	Level IV
Teacher Assessment (T_{ijst})				
Level I	37%	14%	3%	1%
Level II	43%	42%	19%	4%
Level III	19%	42%	64%	43%
Level IV	1%	2%	14%	52%

Table 5: Joint Distribution of A_{ijst} and T_{ijst}: Math

	Achievement Level (A_{ijst})			
	Level I	Level II	Level III	Level IV
Teacher Assessment (T_{ijst})				
Level I	36%	16%	5%	1%
Level II	38%	37%	21%	6%
Level III	25%	41%	56%	38%
Level IV	2%	6%	18%	55%

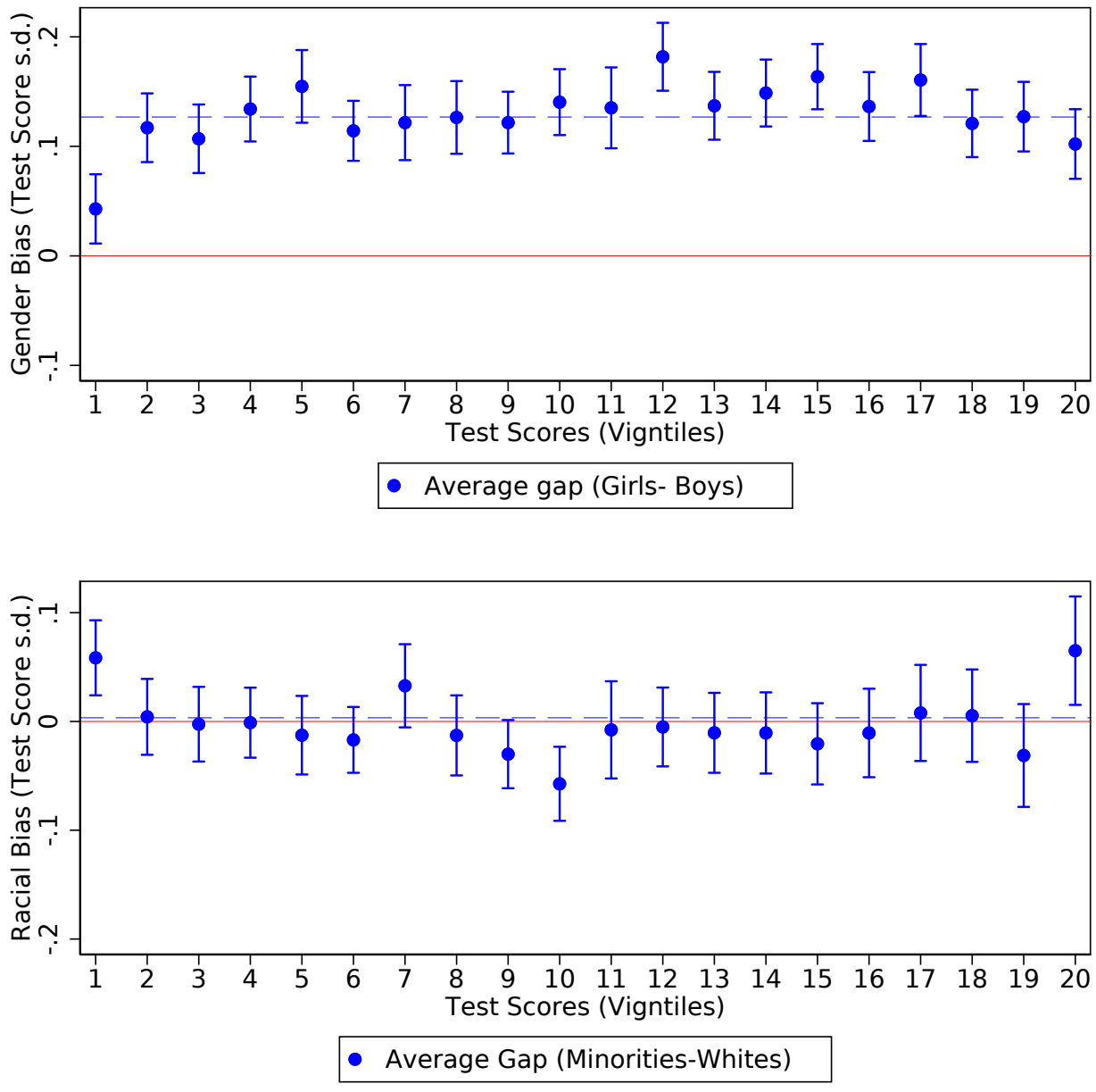
Table 6: Variance of $\hat{\phi}_j$: Math Teachers

	Test Scores	Bias
Total Var: Var($\hat{\phi}_j$)	0.093	0.233
Adjusted Var: Var(ϕ_j)	0.086	0.205

Table 7: Variance of $\hat{\phi}_j$: English Teachers

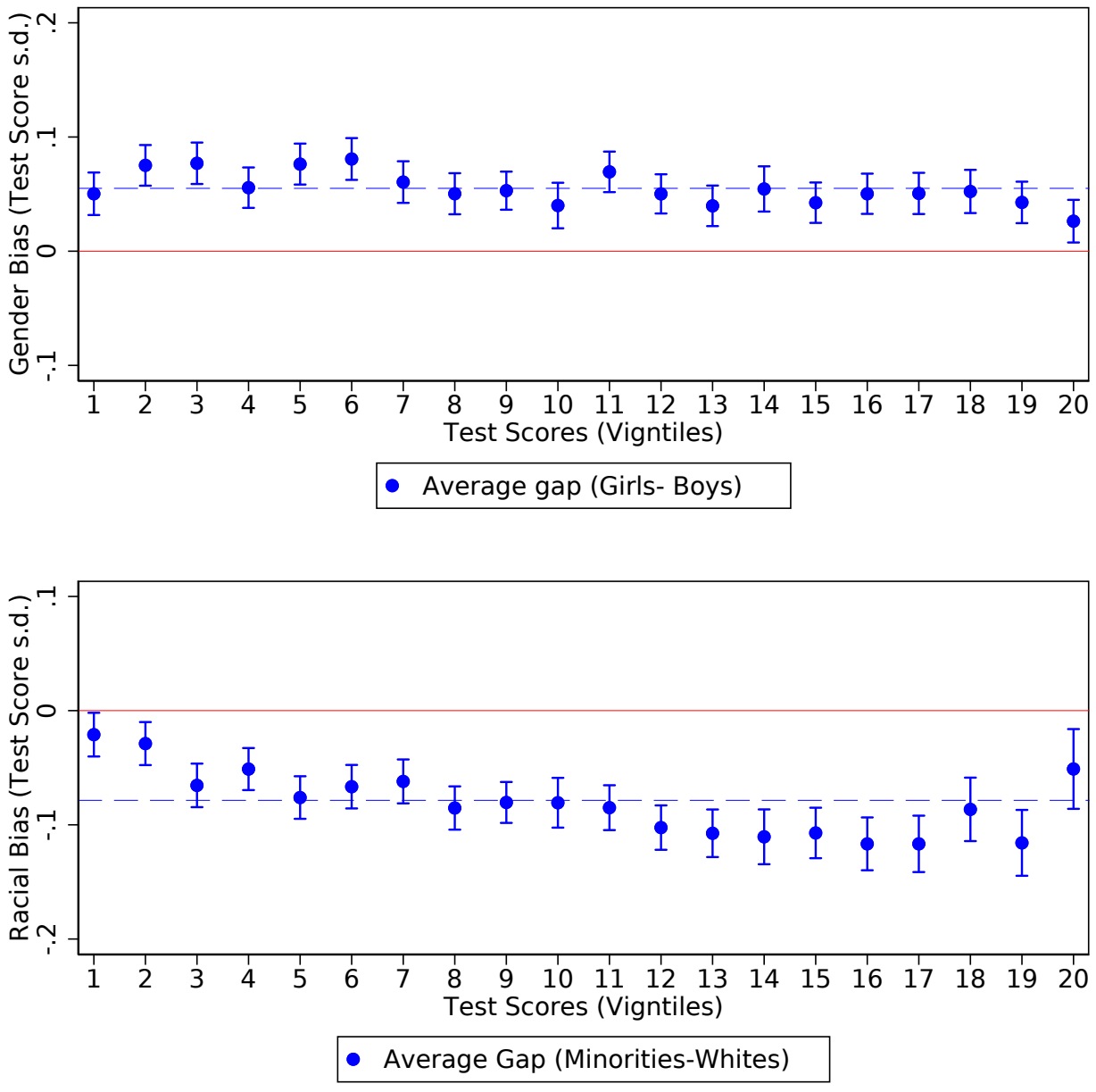
	Test Scores	Bias
Total Var: Var($\hat{\phi}_j$)	0.039	0.123
Adjusted Var: Var(ϕ_j)	0.032	0.109

Figure 3: Unadjusted Differences in Predicted Teacher Assessments: Math



Notes: Each subplot shows the unadjusted differences in assessments (measured as test score standard devations) for each vigntile of the standardized math test score distribution. Each estimate corresponds to the coefficient $\beta + \gamma^v$ in (3.1). This estimation considers the total number of assessments for math courses available in the sample between 2007 and 2013.

Figure 4: Unadjusted Differences in Predicted Teacher Assessments: English



Notes: Each subplot shows the unadjusted differences in assessments (measured as test score standard devations) for each vigntile of the standardized English I test score distribution. Each estimate corresponds to the coefficient $\beta+\gamma^v$ in (3.1). This estimation considers the total number of assessments for English I courses available in the sample between 2007 and 2012.

Figure 5: Number of Assessments Reported by Each Teacher

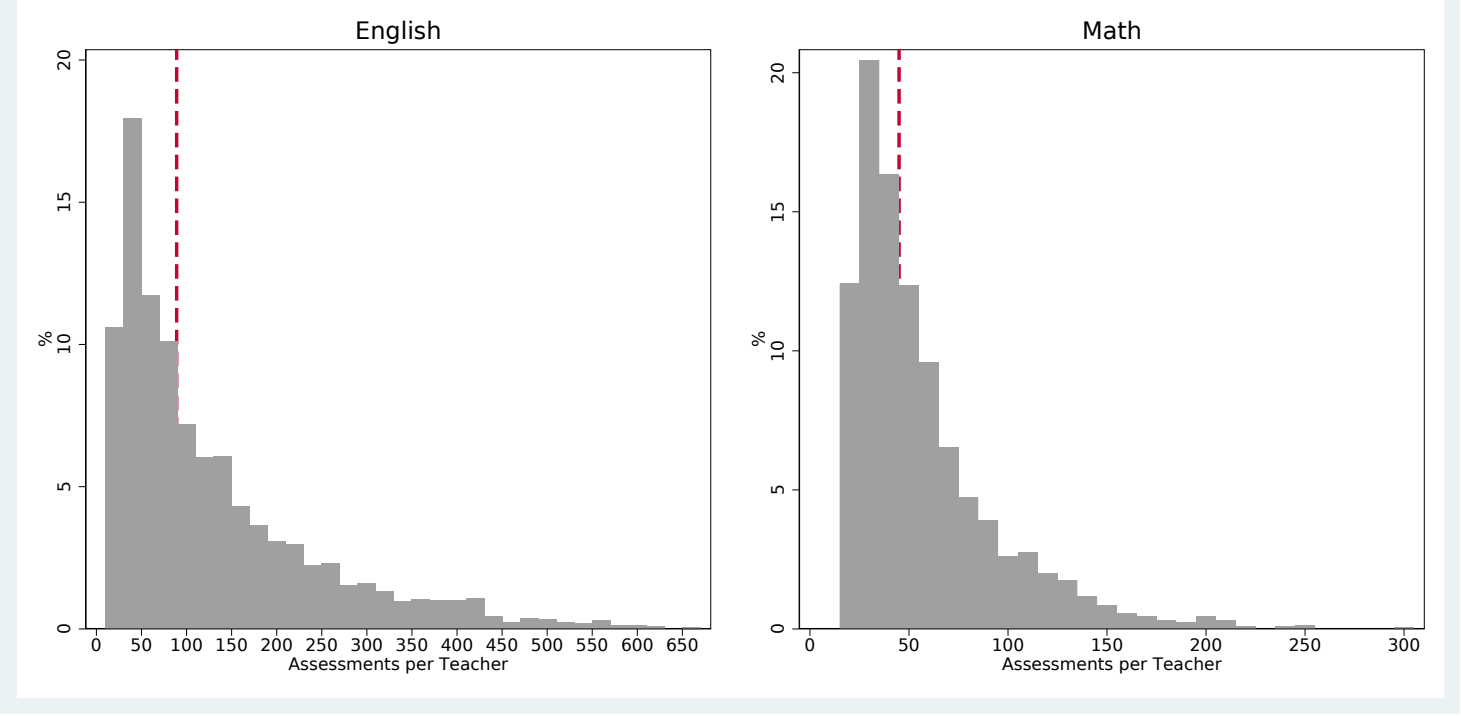


Figure 6: Distribution of Bias by Subject

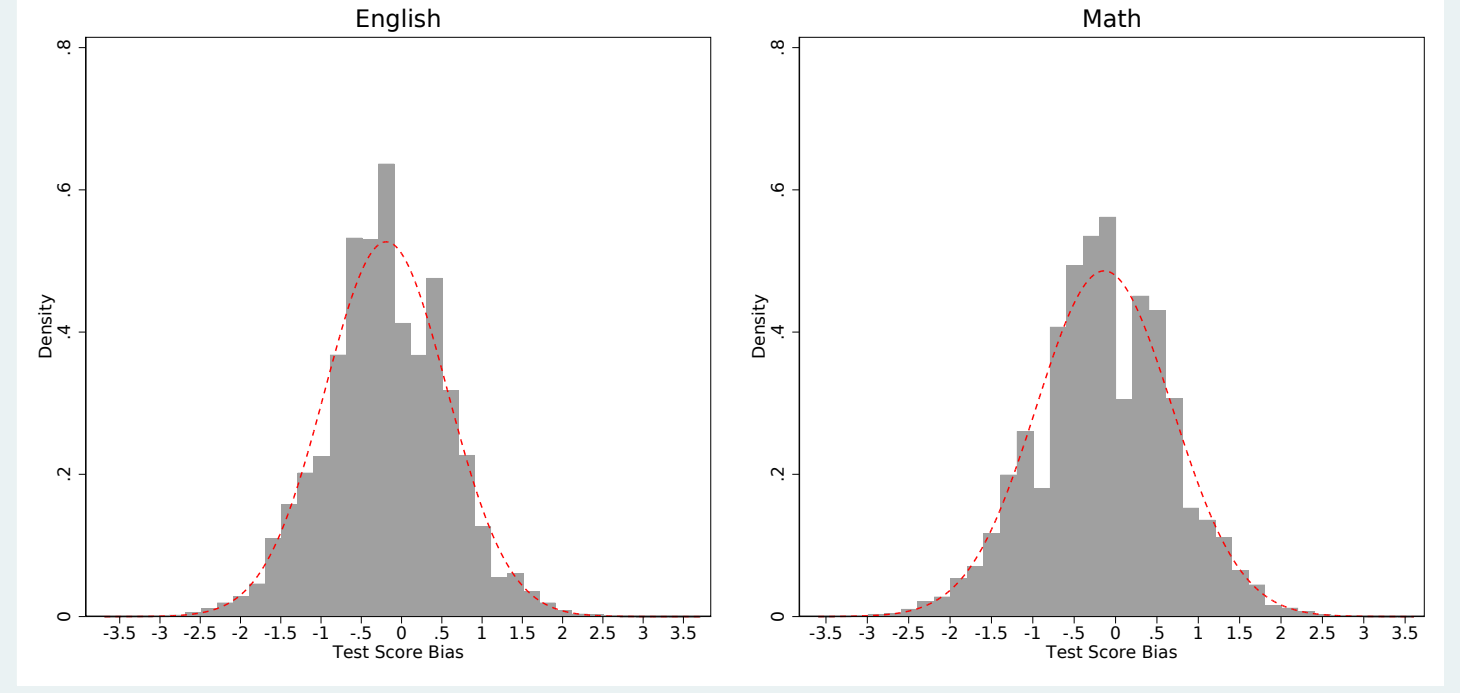
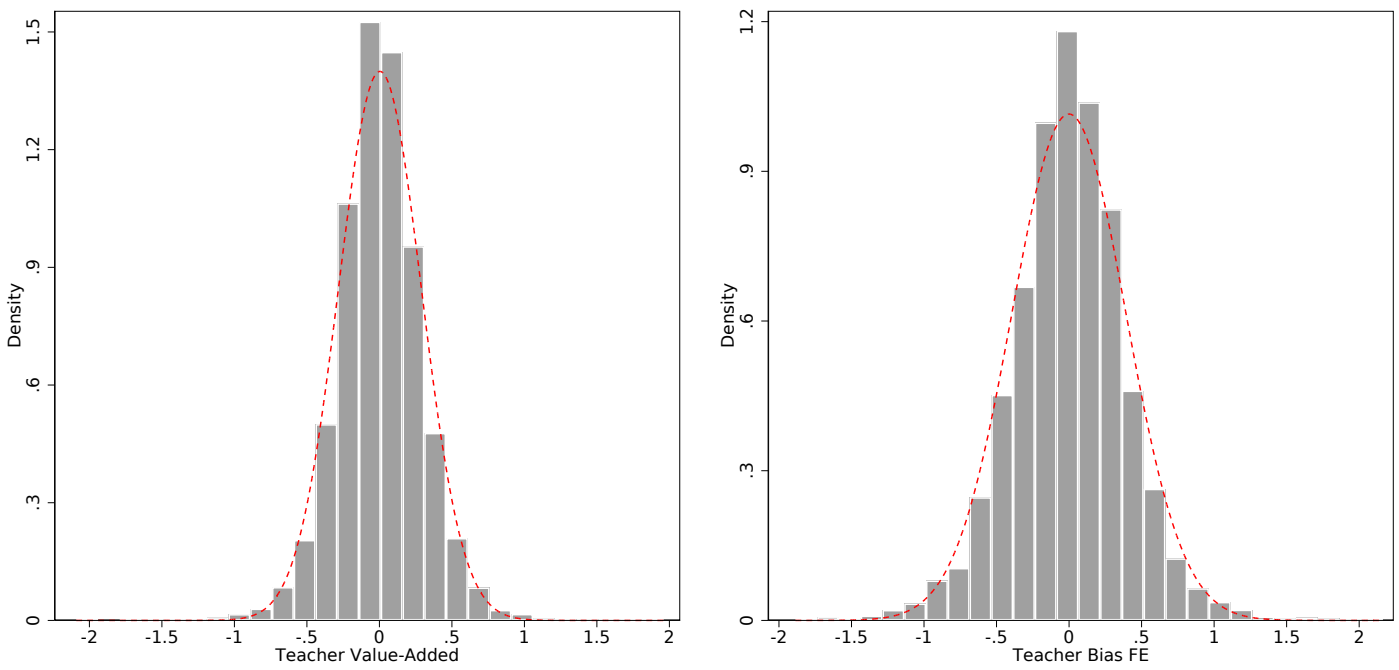


Figure 7: Distribution of the Estimated Teacher FE



(a) Teacher value-added ($\hat{\phi}_j^{VA}$)

(b) Teacher Bias ($\hat{\phi}_j^{B}$)

Notes: This plot shows the raw distribution of the teacher fixed effects $\hat{\phi}_j^{VA}$ and $\hat{\phi}_j^{B}$, estimated using (5.1), separately by subject. The measure of bias corresponds to the difference between the average test score of the students rated in the level assessed by the teacher and the student's test score.

Table 8: Correlation of Teacher Fixed-Effect Estimates

	Test Scores ($\hat{\phi}_j^{VA}$)	Bias ($\hat{\phi}_j^{B}$)
Test Scores ($\hat{\phi}_j^{VA}$)	1	
Bias ($\hat{\phi}_j^{B}$)	-0.340	1

Notes: This matrix reports the correlation between the teacher fixed effects estimated from (6.1), using the pooled leave-year-out estimates. Number of observations: 10,168.

Table 9: Gender and Racial Differences in Assessments - OLS and IV Estimates

	Dependent Variable: $\bar{\theta}^T_{jst} - \theta_{ijst}$				
	All			English I	Algebra I
	(OLS)	(OLS)	(IV)	(IV)	(IV)
Minority	-0.072***	-0.037***	-0.020***	-0.032***	0.021
	(0.003)	(0.006)	(0.006)	(0.007)	(0.014)
Student-Teacher Minority	0.027***	0.022**	0.022**	0.020**	0.022
	(0.009)	(0.009)	(0.009)	(0.010)	(0.016)
Female	0.088***	0.102***	0.088***	0.054***	0.165***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.011)
Subject FE	Yes	Yes	Yes	Yes	Yes
Teacher FE	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes
Student Controls	No	Yes	Yes	Yes	Yes
Order of polynomial on 9th grade scores	3rd	3rd	3rd	3rd	3rd
Observations	455403	405132	405132	311299	93833
R^2	0.38	0.42	0.12	0.13	0.10

Notes: Each column shows the result of regressing the student-level bias on the student's minority and female status, an indicator equals to one if the student and the teacher belong to the minority group, and the additional controls indicated. Each regression includes classroom and teacher fixed effects, as well as a third-order degree polynomial in the contemporaneous test score obtained by the student in 9th grade. Controls include a third degree polynomial on the English and math student's test scores in 8th and 7th grades, number of suspensions, absences, and repeater status in 8th and 7th grades. Standard errors are clustered at the teacher level. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 10: Gender and Racial Differences in Assessments - OLS and IV Estimates

	Bias $(\bar{\theta}^T_{jst} - \theta_{ijst})$		Under-assess $(\mathbb{1}\{T_{ijst} < A_{ijst}\})$		Over-assess $(\mathbb{1}\{T_{ijst} > A_{ijst}\})$	
	(OLS)	(IV)	(OLS)	(IV)	(OLS)	(IV)
Minority	-0.037***	-0.020***	0.018***	0.009***	-0.015***	-0.009***
	(0.006)	(0.006)	(0.004)	(0.004)	(0.003)	(0.003)
Student-Teacher Minority	0.022***	0.022**	-0.013**	-0.012**	0.002	0.003
	(0.009)	(0.009)	(0.005)	(0.005)	(0.005)	(0.005)
Female	0.102***	0.088***	-0.053***	-0.045***	0.031***	0.027***
	(0.005)	(0.006)	(0.003)	(0.003)	(0.003)	(0.003)
Subject FE	Yes	Yes	Yes	Yes	Yes	Yes
Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes
Student Controls	Yes	Yes	Yes	Yes	Yes	Yes
Order of polynomial on 9th grade scores	3rd	3rd	3rd	3rd	3rd	3rd
Observations	405132	405132	405132	405132	405132	405132
R^2	0.42	0.12	0.29	0.05	0.32	0.09

Notes: Each column shows the result of regressing the corresponding dependent variable on the student's minority and female status, an indicator equals to one if the student and the teacher belong to the minority group, and the additional controls indicated. Each regression includes classroom and teacher fixed effects, as well as a third-order degree polynomial in the contemporaneous test score obtained by the student in 9th grade. Student controls include a third degree polynomial on the English and math student's test scores in 8th and 7th grades, number of suspensions, absences, and repeater status in 8th and 7th grades. Standard errors are clustered at the teacher level. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 11: Correlation of Teacher Fixed Effects with Observable Characteristics

	Dep. Var.: $\phi_j^B < 0$			Dep. Var.: $\phi_j^B > 0$		
	(1)	(2)	(3)	(4)	(5)	(6)
Value-Added	0.085***		0.070***	0.004		0.004
	(0.014)		(0.017)	(0.017)		(0.019)
Female		0.018	0.010		-0.055*	-0.055*
		(0.032)	(0.032)		(0.032)	(0.032)
White		-0.128	-0.103		-0.051	-0.051
		(0.134)	(0.135)		(0.123)	(0.123)
Black		-0.133	-0.100		0.035	0.036
		(0.138)	(0.139)		(0.128)	(0.127)
Hispanic		-0.086	-0.057		-0.179	-0.178
		(0.196)	(0.195)		(0.159)	(0.158)
Asian		-0.248	-0.197		0.307	0.308
		(0.254)	(0.245)		(0.254)	(0.254)
Master's degree		-0.023	-0.024		-0.004	-0.004
		(0.028)	(0.028)		(0.026)	(0.026)
Other advanced degree		-0.228*	-0.214*		0.000	0.001
		(0.130)	(0.130)		(0.122)	(0.122)
Subject FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2441	2422	2422	2509	2495	2495
R^2	0.01	0.01	0.01	0.01	0.01	0.01

Each column displays the estimates from a regression of the teacher fixed effect onto the teacher value-added and time-invariant teacher characteristics, using the pooled sample of teachers across subjects. The dependent variable corresponds to the estimated teacher fixed effect in equation (5.1). Value-added corresponds to the estimated test score value-added for each teacher (ϕ_j^T).
*, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 12: Estimates for Experience Profile

	Estimates for $\hat{f}(exp_{jt})$		
		Alternative Dep. Variable	
	Bias $(\bar{\theta}^{T}_{jst} - \theta_{ijst})$	Under-assess $(\mathbb{1}\{T_{ijst} < A_{ijst}\})$	Over-assess $(\mathbb{1}\{T_{ijst} > A_{ijst}\})$
	(1)	(2)	(3)
Teacher Experience:			
Years $\in \{1, 2\}$	-0.014	0.004	-0.011**
	(0.017)	(0.009)	(0.007)
Years $\in \{3, 5\}$	-0.028	0.003	-0.023***
	(0.022)	(0.012)	(0.009)
Years $\in \{6, 10\}$	-0.053*	0.015	-0.035***
	(0.029)	(0.016)	(0.012)
Years $\in \{11, 20\}$	-0.045	0.005	-0.045***
	(0.036)	(0.020)	(0.014)
Years ≥ 21	-0.047	0.004	-0.047**
	(0.048)	(0.027)	(0.019)
Subject FE	Yes	Yes	Yes
Teacher FE	Yes	Yes	Yes
School FE	Yes	Yes	Yes
Student Controls	Yes	Yes	Yes
Order of polynomial on 8th grade scores	3rd	3rd	3rd
Observations	407226	407225	407225
R^2	0.31	0.15	0.17

Notes: Each regression show the estimates of $\hat{f}(exp_{jt})$ in equation (5.1) using the pooled sample across subjects. Each regression includes teacher fixed effects, school fixed effects, and subject fixed effects. Zero years of experience corresponds to the omitted category. Clustered standard errors at the teacher level displayed in parentheses.*, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 13: Outcomes in 9th-12th Grades: Main Specification

	Contemporaneous (9th Grade)		12th Grade			
	Test Score	Plans to Attend College	GPA	Plans to Attend College	SAT Taker	SAT Score
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized Teacher FE:						
Bias x Minority	0.0056**	-0.0028	0.0109***	-0.0014	0.0038**	2.0212***
	(0.0028)	(0.0019)	(0.0031)	(0.0017)	(0.0020)	(0.6178)
Bias x Female	0.0016	0.0078***	0.0045**	0.0054***	0.0034**	1.0545**
	(0.0024)	(0.0017)	(0.0023)	(0.0016)	(0.0015)	(0.4555)
Bias (ϕ_j^B)	-0.0094***	-0.0028*	-0.0035	-0.0011	-0.0017	-1.4218***
	(0.0031)	(0.0015)	(0.0022)	(0.0014)	(0.0014)	(0.4609)
Value-Added (ϕ_j^{VA})	0.0484***	0.0029***	0.0033*	0.0003	0.0019*	0.3158
	(0.0031)	(0.0010)	(0.0017)	(0.0010)	(0.0011)	(0.3292)
Dep. Var. Mean	0.2134	0.7797	3.1818	0.8574	0.4836	1000.36
School-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	320957	290636	251577	246778	321841	151804
R^2	0.68	0.13	0.71	0.13	0.36	0.80

Notes: Clustered standard errors at the teacher level in parentheses. Each regression also includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and reading test scores in 8th grade and 7th grade, number of honors classes taken in 9th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and one-leave-out classroom averages (share of students by race and gender, average math and read scores in 8th grade, share of students by free lunch and reduced-price lunch status, average number of absences, suspensions in 8th grade and 7th grade). *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 14: Test for Student Sorting Using 8th Grade Test Scores

	Math 8th Grade			Read 8th Grade		
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized Teacher FE						
Test Score ($\hat{\phi}_j^{VA}$)	-0.0002		-0.0005	0.0000		-0.0002
	(0.0002)		(0.0004)	(0.0002)		(0.0003)
Bias ($\hat{\phi}_j^{B}$)		-0.0000	-0.0003		0.0003	0.0001
		(0.0003)	(0.0003)		(0.0003)	(0.0003)
School-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	580021	398352	368796	580021	398352	368796
R^2	0.75	0.75	0.75	0.70	0.70	0.70

Clustered standard errors at the teacher level in parentheses

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

Table 15: Test for Student Sorting Using 8th Grade Behaviors

	Days Absent 8th Grade			Suspended 8th Grade		
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized Teacher FE						
Test Score ($\hat{\phi}_j^{VA}$)	-0.0022		-0.0016	-0.0000		-0.0000
	(0.0051)		(0.0077)	(0.0001)		(0.0002)
Bias ($\hat{\phi}_j^{B}$)		0.0079	0.0104		0.0001	0.0002
		(0.0068)	(0.0075)		(0.0001)	(0.0002)
School-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	577082	396322	366961	580021	398352	368796
R^2	0.42	0.42	0.42	0.15	0.15	0.15

Clustered standard errors at the teacher level in parentheses

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

Table 16: Outcomes in 9th-12th Grades: Using Within-School Across-Cohort Variation

	Contemporaneous (9th Grade)		12th Grade			
	Test Score	Plans to Attend College	GPA	Plans to Attend College	SAT Taker	SAT Score
	(1)	(2)	(3)	(4)	(5)	(6)
Mean Standardized FE:						
Bias x Minority	0.0018	-0.0010	0.0101***	-0.0022	0.0022	2.2689***
	(0.0038)	(0.0020)	(0.0038)	(0.0019)	(0.0022)	(0.6663)
Bias x Female	0.0002	0.0077***	0.0060**	0.0082***	0.0030**	1.5417**
	(0.0028)	(0.0019)	(0.0025)	(0.0018)	(0.0015)	(0.4875)
Bias (ϕ_j^B)	-0.0152***	-0.0042*	-0.0060**	-0.0033*	-0.0026	-1.7466***
	(0.0042)	(0.0018)	(0.0029)	(0.0017)	(0.0017)	(0.5361)
Value-Added (ϕ_j^{VA})	0.0350***	0.0032***	0.0020	0.0030**	0.0034**	0.1449
	(0.0039)	(0.0012)	(0.0025)	(0.0014)	(0.0015)	(0.4076)
Dep. Var. Mean	0.2134	0.7797	3.1818	0.8574	0.4836	1000.36
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	321623	291207	252184	247362	322509	152220
R^2	0.67	0.13	0.70	0.11	0.36	0.79

Notes: Clustered standard errors at the teacher level in parentheses. Each regression also includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and reading test scores in 8th grade and 7th grade, number of honors classes taken in 9th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and one-leave-out classroom averages (share of students by race and gender, average math and read scores in 8th grade, share of students by free lunch and reduced-price lunch status, average number of absences, suspensions in 8th grade and 7th grade. *, **, *** denote significance at the 10%, 5%, 1% levels, respectively.

Table 17: Outcomes in 9th-12th Grades: Using Split-Sample IV

	Contemporaneous (9th Grade)		12th Grade			
	Test Score	Plans to Attend College	GPA	Plans to Attend College	SAT Taker	SAT Score
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized Teacher FE:						
Bias x Minority	0.0224	0.0055	0.0920***	-0.0216	0.0381***	3.5467***
	(0.0151)	(0.0114)	(0.0236)	(0.0135)	(0.0134)	(1.2136)
Bias x Female	0.0148	0.0267***	0.0144	0.0500***	0.0178**	2.0571**
	(0.0122)	(0.0096)	(0.0127)	(0.0117)	(0.0087)	(0.8567)
Bias (ϕ_j^B)	-0.0145	0.0093	0.0085	-0.0082	-0.0008	-1.9801***
	(0.0243)	(0.0113)	(0.0196)	(0.0112)	(0.0114)	(0.6951)
Value-Added (ϕ_j^{VA})	0.0663***	0.0086**	0.0204***	0.0046	0.0092**	0.9891
	(0.0090)	(0.0043)	(0.0075)	(0.0038)	(0.0044)	(0.8763)
Dep. Var. Mean	0.2441	0.7758	3.1659	0.8562	0.4770	996.41
School-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	205541	190685	163052	157439	205808	96447
R^2	0.21	0.08	0.25	0.09	0.18	0.24

Notes: For each regression I compute the teacher fixed effects in even and odd years separately and use estimates in even years as an instrument for estimates in odd years, and vice versa. Each regression also includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and reading test scores in 8th grade and 7th grade, number of honors classes taken in 9th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and one-leave-out classroom averages (share of students by race and gender, average math and read scores in 8th grade, share of students by free lunch and reduced-price lunch status, average number of absences, suspensions in 8th grade and 7th grade. Clustered standard errors at the teacher level in parentheses. *, **, *** denote significance at the 10%, 5%, 1% levels, respectively.

A Appendix

A.1 Using Alternative Measures of Mis-Assessment

As a robustness check, I compute teacher fixed effects using each of the alternative measures defined at the end of section 5.2. Recall that these variables were defined as:

Under-assessment:

$$\mathbb{1}\{T_{ijst} < A_{ijst}\}$$

Over-assessment:

$$\mathbb{1}\{T_{ijst} > A_{ijst}\}$$

Precision:

$$|\overline{\theta}^{T}_{jst} - \theta_{ijst}|$$

I employ a similar specification to (5.1) to estimate each teacher fixed effect, but instead of using the variable *bias* ($\overline{\theta}^{T}_{jst} - \theta_{ijst}$) as the dependent variable, I consider each of the variables above. Let ϕ^{U}_{j}, ϕ^{O}_{j}, and ϕ^{P}_{j} denote the teacher fixed effects associated to *under-assessment*, *over-assessment*, and *precision*, respectively.

Figure A1 shows the distribution of the estimated fixed effects in each case. Table A6 shows the correlation of the teacher bias fixed effects with the estimates obtained using these alternative measures. This table shows that the teacher fixed effects associated with under-assessment and over-assessment have a high positive correlation with the measure of bias employed in the analysis of section 6.4. On the other hand, precision is less correlated with bias since it quantifies differences in absolute value. The negative correlation between ϕ^{B}_{j} and ϕ^{P}_{j} suggests that teachers who are more likely to over-assess students tend to mis-assess by a smaller amount than teachers who are more likely to under-assess.

Table A7 displays the raw and adjusted variances for each distribution, separately by subject. For under-assessment, the adjusted variance corresponds to 0.059 and 0.036 for math and English teachers, respectively. Alternatively, an increase of 1 s.d. in $\hat{\phi}^{U}_{j}$ associates with an increase of 24 and 19 percentage points in the probability of under-assessing all students. Similar numbers follow when considering an increase of 1 s.d. in $\hat{\phi}^{O}_{j}$. Finally, since $\hat{\phi}^{P}_{j}$ corresponds to the teacher-specific component of the absolute differences between the test score and the teacher assessment, an increase of 1 s.d. in $\hat{\phi}^{P}_{j}$ associates with an increase in mis-assessment of 0.3 test score standard deviations, regardless of the mis-assessment's sign.

Tables A8, A9, and A10 replicate the main analysis, using each alternative measure of mis-assessing to characterize teacher behavior. All these analyses use the specification (5.4) and control for

teacher test score value-added. Tables A8 and A9 display the results after using under-assessment and over-assessment, respectively. The heterogeneity patterns for girls and minorities are similar to the ones discussed in the main analysis. Conditional on being exposed to a teacher more likely to under-assess or over-assess, girls and minorities are more affected. This result is largely consistent with the high correlation between the teacher fixed effects shown in Table A6. Simultaneously, the use of precision informs about the extent to which being excessively mis-assessed can negatively affect outcomes. Table A10 shows that an increase of 1 s.d. in $\hat{\phi}^P$, equivalent to *mis-assess* students by 0.3 test score standard deviations, leads to decreases in all the outcomes for girls minorities, except for contemporaneous test scores and SAT scores.

A.2 Additional Figures and Tables

Table A1: Gender and Racial Differences in Assessments - OLS and IV Estimates

	Bias $(\bar{\theta}^T_{jst} - \theta_{ijst})$		Under-assess $(\mathbb{1}\{T_{ijst} < A_{ijst}\})$		Over-assess $(\mathbb{1}\{T_{ijst} > A_{ijst}\})$	
	(OLS)	(IV)	(OLS)	(IV)	(OLS)	(IV)
Minority	-0.037***	-0.020***	0.018***	0.009***	-0.015***	-0.009***
	(0.006)	(0.006)	(0.004)	(0.004)	(0.003)	(0.003)
Student-Teacher Minority	0.022***	0.022**	-0.013**	-0.012**	0.002	0.003
	(0.009)	(0.009)	(0.005)	(0.005)	(0.005)	(0.005)
Female	0.102***	0.088***	-0.053***	-0.045***	0.031***	0.027***
	(0.005)	(0.006)	(0.003)	(0.003)	(0.003)	(0.003)
Subject FE	Yes	Yes	Yes	Yes	Yes	Yes
Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes
Student Controls	Yes	Yes	Yes	Yes	Yes	Yes
Order of polynomial on 9th grade scores	3rd	3rd	3rd	3rd	3rd	3rd
Observations	405132	405132	405132	405132	405132	405132
R^2	0.42	0.12	0.29	0.05	0.32	0.09

Notes: Each column shows the result of regressing the corresponding dependent variable on the student's minority and female status, an indicator equals to one if the student and the teacher belong to the minority group, and the additional controls indicated. Each regression includes classroom and teacher fixed effects, as well as a third-order degree polynomial in the contemporaneous test score obtained by the student in 9th grade. Student controls include a third degree polynomial on the English and math student's test scores in 8th and 7th grades, number of suspensions, absences, and repeater status in 8th and 7th grades. Standard errors are clustered at the teacher level. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A2: Robustness Check: Testing for Miss-classification

	Under-assess $\mathbb{1}\{T_{ijst} < A_{ijst}\}$		Over-assess $\mathbb{1}\{T_{ijst} > A_{ijst}\}$	
	Boys	Girls	Boys	Girls
	(1)	(2)	(3)	(4)
Panel A: English				
Minority	0.023***	0.027***	-0.013***	-0.018***
	(0.003)	(0.003)	(0.003)	(0.003)
Student-Teacher Minority	-0.008	-0.015*	0.001	0.001
	(0.009)	(0.008)	(0.009)	(0.009)
Observations	149919	153617	110077	97435
R^2	0.36	0.36	0.42	0.43
Panel B: Math				
Minority	-0.001	0.015**	0.010	0.009
	(0.007)	(0.006)	(0.007)	(0.007)
Student-Teacher Minority	-0.020	-0.012	0.000	-0.009
	(0.016)	(0.013)	(0.016)	(0.017)
Observations	58269	58506	47452	42594
R^2	0.38	0.36	0.43	0.40
9th Grade Test Scores	Yes	Yes	Yes	Yes
Teacher FE	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes

Notes: This table shows the estimates of equation (6.1) restricting the sample to observations where it is possible to distinguish whether each teacher mis-assesses students (Levels II-IV for under-assessment and Levels I-III for over-assessment). Each column shows the result of regressing the corresponding indicator variable on the student's minority status, an indicator equals to one if the student and the teacher belong to the minority group, and the additional controls indicated. 9th Grade Test Scores includes a third degree polynomial in the math test score obtained by the student in 9th grade. Controls include a third degree polynomial on the English and math student's test scores in 8th and 7th grades, number of suspensions, absences, and repeater status in 8th and 7th grades, and one-leave-out classroom average characteristics (share of students by race and gender, average math and reading scores in 8th grade, share of students by free lunch and reduced-price lunch status, average number of absences, average number of suspensions in 8th grade and 7th grade). Standard errors are clustered at the teacher level. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A3: Estimates for Experience, by Teacher Gender

	Estimates for $\hat{f}(exp_{jt})$			
	Under-assess $\mathbb{1}\{T_{ijst} < A_{ijst}\}$		Over-assess $\mathbb{1}\{T_{ijst} > A_{ijst}\}$	
	Female	Male	Female	Male
Exp $\in [1, 2]$	0.008	-0.014**	-0.022***	-0.017
	(0.010)	(0.017)	(0.007)	(0.012)
Exp $\in [3, 5]$	0.003	-0.001	-0.032***	-0.037**
	(0.012)	(0.024)	(0.010)	(0.016)
Exp $\in [6, 10]$	0.015	0.042	-0.043***	-0.068***
	(0.016)	(0.034)	(0.012)	(0.021)
Exp $\in [11, 20]$	0.010	0.061	-0.053***	-0.078***
	(0.018)	(0.045)	(0.014)	(0.026)
Exp >20	-0.073	0.054	-0.051***	-0.055
	(0.025)	(0.065)	(0.017)	(0.046)
School-Year FE	Yes	Yes	Yes	Yes
Teacher FE	Yes	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes	Yes
Observations	353,668	88,928	353,668	88,928
R^2	0.24	0.27	0.29	0.33

Each column displays the coefficients $\hat{f}(exp_{jt})$ associated to regression (6.1), where the dependent variable is an indicator of under-assessment and an indicator of over-assessment, separately by teacher gender. The omitted category corresponds to zero years of experience. Clustered standard errors at the teacher level displayed in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A4: Estimates for Experience, by Teacher Race

	Estimates for $\hat{f}(exp_{jt})$			
	Under-assess $\mathbb{1}\{T_{ijst} < A_{ijst}\}$		Over-assess $\mathbb{1}\{T_{ijst} > A_{ijst}\}$	
	White	Minority	White	Minority
Exp $\in [1, 2]$	-0.000	-0.002	-0.017***	-0.017
	(0.009)	(0.022)	(0.006)	(0.021)
Exp $\in [3, 5]$	-0.008	0.016	-0.025***	-0.048**
	(0.011)	(0.032)	(0.008)	(0.024)
Exp $\in [6, 10]$	0.006	0.047	-0.037***	-0.078**
	(0.015)	(0.043)	(0.010)	(0.034)
Exp $\in [11, 20]$	0.003	0.061	-0.048***	-0.090**
	(0.018)	(0.065)	(0.012)	(0.041)
Exp >20	-0.011	0.013	-0.048***	-0.085*
	(0.024)	(0.088)	(0.016)	(0.046)
School-Year FE	Yes	Yes	Yes	Yes
Teacher FE	Yes	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes	Yes
Observations	384,771	53,846	384,771	53,846
R^2	0.23	0.29	0.28	0.35

Each column displays the coefficients $\hat{f}(exp_{jt})$ associated to regression (6.1), where the dependent variable is an indicator of under-assessment and an indicator of over-assessment, separately by teacher race. Minority denotes black and Hispanic teachers, according to the teacher's race classification. The omitted category corresponds to zero years of experience. Clustered standard errors at the teacher level displayed in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Correlation of Teacher Fixed Effects with Observable Characteristics

	Dependent Variable: Teacher Value-Added ($\hat{\phi}_j^{VA}$)					
	All Teachers		By Gender		By Race	
			Female	Male	White	Non-White
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher Bias (1st quintile)	-0.155***	-0.100***	-0.123***	-0.025	-0.082*	-0.198**
	(0.045)	(0.041)	(0.048)	(0.082)	(0.045)	(0.100)
Teacher Bias (2nd quintile)	0.023	0.022	0.040	-0.033	0.044	-0.099
	(0.044)	(0.039)	(0.039)	(0.044)	(0.080)	(0.042)
Teacher Bias (4th quintile)	0.027	0.022	0.002	0.103	0.044	-0.097
	(0.042)	(0.038)	(0.043)	(0.082)	(0.042)	(0.087)
Teacher Bias (5th quintile)	0.040	0.036	0.032	0.054	0.080*	-0.167*
	(0.045)	(0.041)	(0.047)	(0.086)	(0.045)	(0.099)
Teacher Controls	No	Yes	Yes	Yes	Yes	Yes
Subject FE	No	Yes	Yes	Yes	Yes	Yes
Observations	4950	4917	3733	1184	4089	828
R^2	0.01	0.16	0.16	0.18	0.16	0.19

Each column displays the estimates from a regression of the teacher value-added onto the teacher bias estimates, by quintiles, and time-invariant teacher characteristics. The omitted category is the third quintile of the teacher bias distribution. Each regression employs the pooled sample of teachers across subjects. Teacher bias corresponds to the estimated teacher fixed effect in equation (5.1), while the dependent variable corresponds to the estimated test score value-added for each teacher (ϕ_j^T). *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A6: Correlation of Teacher Fixed-Effect Estimates

	Bias ($\hat{\phi}_j^B$)	Under-assessment ($\hat{\phi}_j^U$)	Over-assessment ($\hat{\phi}_j^O$)	Precision ($\hat{\phi}_j^P$)
Bias ($\hat{\phi}_j^B$)	1			
Under-assessment ($\hat{\phi}_j^U$)	-0.848	1		
Over-assessment ($\hat{\phi}_j^O$)	0.825	-0.903	1	
Precision ($\hat{\phi}_j^P$)	-0.121	0.114	0.024	1

Notes: This matrix reports the correlation between the teacher fixed effects estimated from (6.1), using the pooled set of leave-year-out estimates.

Table A7: Variance of $\hat{\phi}_j$ by Subject

	Math Teachers	English Teachers
Under-assessment (ϕ_j^U):		
Total Var	0.066	0.040
Adjusted Var	0.059	0.036
Over-assessment (ϕ_j^O):		
Total Var	0.068	0.038
Adjusted Var	0.063	0.033
Precision (ϕ_j^P):		
Total Var	0.099	0.041
Adjusted Var	0.086	0.037

Table A8: Outcomes in 9th-12th Grades: Using Under-assessment FE

	Contemporaneous (9th Grade)		12th Grade			
	Test Score	Plans to Attend College	GPA	Plans to Attend College	SAT Taker	SAT Score
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized Teacher FE:						
Under-assess x Minority	-0.0046*	0.0012	-0.0101***	0.0012	-0.0047**	-2.0493***
	(0.0027)	(0.0018)	(0.0031)	(0.0017)	(0.0019)	(0.5692)
Under-assess x Female	-0.0026	-0.0061***	-0.0020	-0.0048***	-0.0041***	-1.2179***
	(0.0023)	(0.0016)	(0.0022)	(0.0015)	(0.0015)	(0.4530)
Under-assess (ϕ_j^U)	0.0075**	0.0029**	0.0025	0.0005	0.0027*	1.1020**
	(0.0030)	(0.0014)	(0.0021)	(0.0014)	(0.0015)	(0.4434)
Value-Added (ϕ_j^{VA})	0.0482***	0.0015	0.0025	0.0002	0.0006	0.5199
	(0.0031)	(0.0010)	(0.0016)	(0.0010)	(0.0011)	(0.3392)
Dep. Var. Mean	0.2134	0.7797	3.1818	0.8574	0.4836	1000.36
School-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	324431	294850	255240	249991	325320	153890
R^2	0.68	0.13	0.71	0.13	0.36	0.80

Notes: Clustered standard errors at the teacher level in parentheses. Under-assess corresponds to the teacher fixed effect computed using (5.1) using $\mathbb{1}\{T_{ijst} < A_{ijst}\}$ as the dependent variable. Each regression also includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and reading test scores in 8th grade and 7th grade, number of honors classes taken in 9th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and one-leave-out classroom averages (share of students by race and gender, average math and read scores in 8th grade, share of students by free lunch and reduced-price lunch status, average number of absences, suspensions in 8th grade and 7th grade). *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A9: Outcomes in 9th-12th Grades: Using Over-assessment FE

	Contemporaneous (9th Grade)		12th Grade			
	Test Score	Plans to Attend College	GPA	Plans to Attend College	SAT Taker	SAT Score
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized Teacher FE:						
Over-assess x Minority	0.0063**	-0.0009	0.0087***	-0.0008	0.0038*	1.1372**
	(0.0026)	(0.0018)	(0.0030)	(0.0017)	(0.0019)	(0.5618)
Over-assess x Female	0.0003	0.0062***	0.0038*	0.0056***	0.0030**	1.2125***
	(0.0023)	(0.0016)	(0.0022)	(0.0015)	(0.0015)	(0.4506)
Over-assess (ϕ_j^O)	-0.0077**	-0.0030**	-0.0008	-0.0012	-0.0018	-1.3246***
	(0.0030)	(0.0014)	(0.0021)	(0.0013)	(0.0015)	(0.4281)
Value-Added (ϕ_j^{VA})	0.0499***	0.0025**	0.0046***	0.0012	0.0022**	0.5052
	(0.0033)	(0.0010)	(0.0017)	(0.0010)	(0.0011)	(0.3268)
Dep. Var. Mean	0.2134	0.7797	3.1818	0.8574	0.4836	1000.36
School-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	323466	293767	253735	248547	324322	153252
R^2	0.68	0.13	0.71	0.13	0.36	0.80

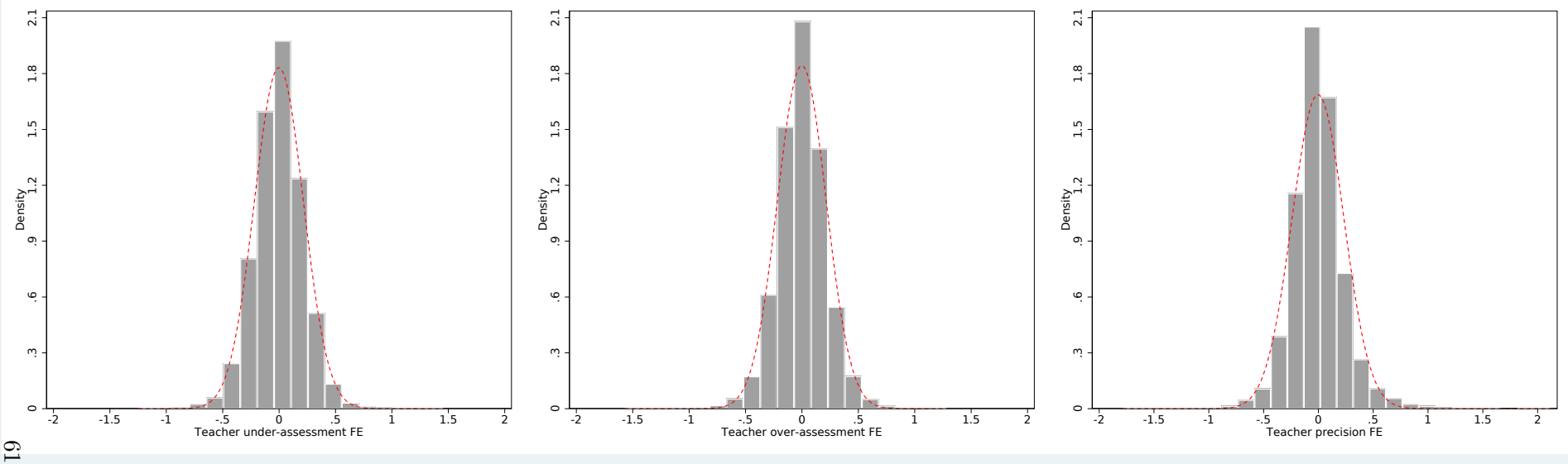
Notes: Clustered standard errors at the teacher level in parentheses. Over-assess corresponds to the teacher fixed effect computed using (5.1) using $\mathbb{1}\{T_{ijst} > A_{ijst}\}$ as the dependent variable. Each regression also includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and reading test scores in 8th grade and 7th grade, number of honors classes taken in 9th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and one-leave-out classroom averages (share of students by race and gender, average math and read scores in 8th grade, share of students by free lunch and reduced-price lunch status, average number of absences, suspensions in 8th grade and 7th grade). *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table A10: Outcomes in 9th-12th Grades: Using Precision FE

	Contemporaneous (9th Grade)		12th Grade			
	Test Score	Plans to Attend College	GPA	Plans to Attend College	SAT Taker	SAT Score
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized Teacher FE:						
Precision x Minority	0.0009	0.0014	-0.0068**	0.0009	-0.0038**	-0.8576
	(0.0031)	(0.0018)	(0.0029)	(0.0017)	(0.0019)	(0.5633)
Precision x Female	0.0005	-0.0049***	-0.0029	-0.0031**	-0.0013	-0.0900
	(0.0022)	(0.0016)	(0.0020)	(0.0014)	(0.0014)	(0.4731)
Precision (ϕ_j^P)	0.0041	0.0028**	0.0037*	0.0007	0.0008	0.7820*
	(0.0031)	(0.0013)	(0.0021)	(0.0012)	(0.0014)	(0.4586)
Value-Added (ϕ_j^{VA})	0.0488***	0.0029**	0.0035***	0.0015	0.006	0.4223
	(0.0032)	(0.0009)	(0.0015)	(0.0009)	(0.0010)	(0.3073)
Dep. Var. Mean	0.2134	0.7797	3.1818	0.8574	0.4836	1000.36
School-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	325039	294558	255191	251133	325895	153223
R^2	0.68	0.13	0.71	0.13	0.36	0.80

Notes: Clustered standard errors at the teacher level in parentheses. Precision corresponds to the teacher fixed effect computed using (5.1) using $|\overline{\theta}^T_{jst} - \theta_{ijst}|$ as the dependent variable. Each regression also includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and reading test scores in 8th grade and 7th grade, number of honors classes taken in 9th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and one-leave-out classroom averages (share of students by race and gender, average math and read scores in 8th grade, share of students by free lunch and reduced-price lunch status, average number of absences, suspensions in 8th grade and 7th grade). *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Figure A1: Alternative Measures of Mis-assessment



(a) Under-assessment ($\hat{\phi}_j^U$) (b) Over-assessment ($\hat{\phi}_j^O$) (b) Precision ($\hat{\phi}_j^P$)

Notes: This plot shows the raw distribution of the teacher fixed effects $\hat{\phi}_j^U$, $\hat{\phi}_j^O$, and $\hat{\phi}_j^P$, respectively, estimated using (5.1), separately by subject. The definitions of under-assessment, over-assessment, and precision corresponds to the ones included at the end of section 5.2.