Teacher Mis-Assessments and High School Outcomes: Heterogeneous Effects by Race and Gender[*]

Alexis Orellana[†]

July 14, 2022

Abstract

Does mis-assessment by teachers on subjective evaluations matter for students' educational outcomes? I employ administrative data from North Carolina that contain standardized test scores and teacher assessments for each ninth-grade student to examine whether exposure to a teacher whose judgments differ systematically from students' achievement levels impacts student outcomes. Exposure to teachers who are more likely to overassess students, relative to what test scores signal, increases GPA and college expectations for girls and non-white students. In terms of SAT scores, I find increases for blacks and Hispanics but decreases for Asian students.

JEL Codes: I21, J15, J16, J24

[*] I am indebted to Ronni Pavan, Josh Goodman, and John Singleton for their support throughout this project. I also thank Lisa Kahn, Marcus Winters, Kegon Tan, and Travis Baseler for helpful suggestions and feedback. This paper uses data from the North Carolina Education Research Data Center (NCERDC). I thank Kara Bonneau for her invaluable assistance with providing access to the administrative data.

[†] Wheelock College of Education and Human Development, Boston University. E-mail: borellan@bu.edu

1 Introduction

For decades the concept of teacher quality has been primarily associated with the ability to raise test scores. Indeed, teachers' multidimensional impacts have only recently been the focus of much research (Jackson, 2018; Kraft, 2019; Petek and Pope, 2021). The scope of an effective teacher does not restrict to instruction, since teachers can also influence parental beliefs (Kinsler and Pavan, 2021) and students' effort choices (Mechtenberg, 2009).[1] Therefore, an important but relatively understudied component of teacher effectiveness corresponds to their aptitude to correctly assess students' learning. Yet, a growing number of papers show that teachers do not judge equally all the students they teach. 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; Botelho et al., 2015).

Research from psychology and economics has shown that inaccurate evaluations can affect worker performance and productivity, and this mechanism can be particularly relevant in educational contexts.[2] The psychology literature suggests that, 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. 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.

[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 (2021).

[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.

Taken together, the available evidence in workplace contexts raises the question of whether this mechanism is also present within schools. In particular, whether some groups of students are more affected than others by exposure to mis-assessment and how relevant these effects are. In this paper, I investigate these two questions, exploiting the availability of both standardized and teacher-reported measures of student achievement. Each year, the North Carolina's State Board of Education classifies students in one out of four achievement levels based on their end-of-year test scores. Simultaneously, teachers assess students using the same scale. Drawing on these two measures, I first show that ninth-grade teachers exhibit systematic differences across observationally equivalent students of different gender and ethnicity. Based on these descriptive findings, I study whether teachers who are more likely to differ in the evaluation of their students' achievement (relative to the standardized test score) have a differential impact on girls (compared to boys) and non-white (compared to white) students. To answer this last question, I employ a two-step approach. First, I employ the difference between a given student i's test score and the average test score of the subset of peers rated by the teacher in the same level as i to characterize how teachers judge students over time. Based on this variable, and exploiting the longitudinal nature of my sample, I estimate a persistent component for each teacher, measured in test score standard deviations (s.d.), which I label *inaccuracy*, and construct leave-year-out, empirical Bayes estimates following the teacher value-added literature. Teachers with higher inaccuracy correspond to teachers whose assessments differ more (positively) from the achievement level indicated by test scores. This approach departs from a small, but growing, literature on the effects that teachers with varying degrees of gender favoritism or implicit bias have on girls and boys (Lavy and Sand, 2018; Carlana, 2019; Lavy and Megalokonomou, 2019; Terrier, 2020). With the exception of Carlana (2019), these papers employ a measure of teacher gender bias at the classroom level, defined as the difference between boys' and girls' average gap between the non-blind and the blind score.[3] Nevertheless, this measure of *relative* bias restricts the analysis of heterogeneous impacts across other dimensions. In the second step, I project my teacher-specific estimates onto several outcomes measured between ninth and twelfth grades,

[3] 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.

such as contemporaneous test scores, intention to attend college, GPA, and SAT scores.

My estimation strategy requires the availability of teacher assessments and an additional, comparable 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 in ninth grade included a question where teachers assessed each student's achievement level. This evaluation, reported when students were taking the exam, is comparable to the achievement levels defined by the State Board of Education every year, which are 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 into schools and classrooms. For this reason, my empirical strategy consists in estimating the effect of exposure to a teacher whose assessments deviate to a greater or lesser degree from the (ex-post) observed test scores. To avoid mechanical endogeneity, I use leave-year-out observations to construct teacher estimates of inaccuracy. To account for student sorting, my preferred specification includes controls for individual ability and behavior using lagged proxies, classroom-level average characteristics, as well as school-track and year fixed effects.

My descriptive analysis shows that North Carolina high-school teachers exhibit significant differences at the moment of assessing females and non-white students, relative to other peers with similar observable characteristics. Consistent with findings reported in previous literature using information from blind and non-blind measures of 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; Shi and Zhu, 2021), teachers are more likely to overrate females relative to boys. They are also more llkely to overrate Asian students and to underrate black and Hispanic students, relative to white students.[4] Never-

[4] These papers 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. 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, 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

theless, these patterns vary substantially by subject. In terms of race-ethnicity, I find that English and Biology teachers are more likely to underassess black and Hispanic students. By contrast, Asian students are judged more favorably across all subjects, although these differences are larger for math teachers. I show that these estimates are robust to the inclusion of classroom and teacher fixed effects, as well as to controlling for 8th grade teacher assessments.

Motivated by these descriptive patterns, I apply my estimation strategy to analyze the heterogeneous effects of exposure to this dimension of teacher effectiveness by students' gender and race-ethnicity. I find that teachers whose assessments systematically favor students relative to the performance level associated to test scores (in other words, teachers who tend to overrate students) have a differential positive effect on girls, as well as on black, Hispanic, and Asian students. An increase of 1 s.d. in the teacher inaccuracy distribution has a positive differential impact on contemporaneous test scores and in the probability of planning to attend college on girls relative to boys. In terms of race-ethnicity, I find that a positive effect for Black students in contemporaneous test scores, relative to white students.

These heterogeneous patterns persist for outcomes observed at the end of high school. For girls, I find a positive differential effect on 12th grade GPA, intention to attend college after graduation, and SAT scores. By contrast, the effects for boys are negative or not statistically different from zero at the 10% level. I also find significant differences for each of these outcomes across racial-ethnic groups. While my estimates for white students show that exposure to more inaccurate teachers does not have impact on 12th grade outcomes for them, I find positive effects for black, Hispanic, and Asian students. Nevertheless, this effect is not observed on SAT scores. While exposure to more inaccurate teachers benefits black and Hispanic students by 0.6 points and 1.4 points, respectively, it reduces SAT scores for Asian students by 3.7 points. I quantify the increase in the explanatory power of teacher effects after including my measure of teacher inaccuracy compared to a specification where

same race. Using data from North Carolina, Rangel and Shi (2021) find that elementary teachers are less likely to overassess black students, while Shi and Zhu (2021) show that the presence of Asian students in the classroom exacerbates the white-Black and white-Hispanic assessment gaps.

only test-score value added is used. I find that accounting for this dimension of teacher quality leads to substantial increases in the predictive power of teacher effects, particularly for outcomes observed in 12th grade.

To address concerns related to selection of students into classrooms, I use two different strategies previously considered in the teacher value-added 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 conduct additional robustness checks by using alternative measures of teacher mis-assessment. I redo my analysis employing binary indicators of overassessment and underassessment to characterize teachers. Finally, I also account for the possibility of other-subject teachers influencing the main results, by using a sub-sample of students linked to more than one teacher and employing an additional set of teacher fixed effects. The main results remain qualitatively similar to my baseline analysis providing additional support to my preferred specification.

Overall, these findings are informative about one relatively understudied dimension of teacher effectiveness. In economics, 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).[5] While this set of papers focuses on test scores as the primary measure of productivity, there is evidence that teacher quality involves other dimensions not captured by test scores (Jackson, 2018; Kraft, 2019; Petek and Pope, 2021). I contribute to this literature by studying the capability to provide accurate assessments as one of these additional teacher quality dimensions. More broadly, my results emphasize the limitations of assuming homogeneous teacher effects, particularly for behaviors or outcomes that do not depend exclusively on cognitive skills. As a second contribution, my results also shed light about the channels through which teacher expectations operate (Papageorge et al., 2020;

[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).

Hill and Jones, 2021). Papageorge et al. (2020) show that tenth-grade teachers' expectations increase the probability of completing a four-year college degree, finding that teachers' lower optimism for black students puts them at a disadvantage relative to white students. My results suggest that the positive effects they report for college completion are mediated by increases in human capital accumulation as well as behaviors that positively impact college enrollment, with substantial differences by gender and ethnicity.

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

2 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.

2.1 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.[6] These exams are not graded by teachers, but scores count as 20% of a student's grade in the respective course.

2.2 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.[7] 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.
- 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.[8] Figure 1 shows an example of the question included in the end-of-grade tests in

[6] 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.

[7] 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.)

[8] Hill and Jones (2021) use these reports as a measure of teacher expectations of students' performance.

2011.[9] 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.

Achievement 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.[10] 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.[11]

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.[12] This would be a concern if any student's unobserved characteristic has already influenced

[9] 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.

[10] 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.

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

[12] 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.

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 matching algorithm, similar to the one used by Mansfield (2015) and Jackson (2018).[13] 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.

2.3 Descriptive Analysis

I finalize this section by presenting some descriptive statistics about the final sample used in the estimation. I also present evidence of the assessment gaps by gender and race 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

[13] 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.

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.

Tables 4 - 6 show the distribution of teacher assessments conditional on achievement levels. The diagonal in each table shows the proportion of correct assessments for each achievement level. The correlation between the achievement levels and teacher assessments is high, supporting the assumption that teachers provide informative reports. For example, considering students whose test score corresponds to level III, teachers assess correctly between 64% and 56% of times. Finally, Figure 6 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.

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 each class between 2007 and 2013. As shown in Table 3, on average, teachers predict students' achievement correctly around 50% of times. With the exception of Algebra I, the distribution is not symmetric and it shows the tendency to to underrate students.

To estimate the unconditional gender and racial assessment gaps in terms of test score standard deviations, I consider the following measure. 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 i's test score θ_{ijst}.[14] Then, I estimate regressions of the form:

$$\overline{\theta}_{jst}^{T} - \theta_{ijst} = \alpha_1 + \sum_{d=2}^{10} \alpha^d \theta_{ijst}^d + \beta I_i + \sum_{d=2}^{10} \gamma^d \left(\theta_{ijst}^d \times I_i\right) + \epsilon_{ijst} \tag{2.1}$$

[14] I discuss in more detail the construction of this variable in section 4.

Where $\overline{\theta}^{T}_{jst}$ corresponds to the average score of all students rated by teacher j in the level T_{ijst}. Therefore, the left-hand side measures the deviation between the score associated to each teacher evaluation and the actual score for each student. θ^{d}_{ijst} is an indicator variable for whether student i's test score is in the d-th decile of the test score distribution. $I_i = \{\text{Girl}, \text{Black}, \text{Hispanic}, \text{Asian}\}$ is a vector of indicator variables where each component is equal to one if student i belongs to a specific subgroup. I employ gender and race-ethnicity to classify students. Figures 3 - 5 plot the coefficients $\beta + \gamma^{d}$ by race and gender, separately by subject. The upper-left plot in each figure shows the average differences by gender, while the upper-right and lower plots show the differences between white and black, Hispanic, and Asian students, respectively. The dashed line in each plot shows the average unadjusted gap between the two corresponding groups of students. The comparison of these plots shows that across all subjects teachers overrate girls across the entire distribution of test scores, and these differences are statistically significant at the 5% level. Nevertheless, in terms of race-ethnicity there is heterogeneity in the sign and the magnitude of the differences relative to white students. While math teachers exhibit only a tendency to overassess Asian with respect to white students, English teachers display significant negative differences for blacks and Hispanics. This behavior is also observed for Biology teachers, although in this case the differences tend to concentrate on students in the lower deciles of the test score distribution.

To account for the presence of confounding factors, such as sorting, teacher practices, or student behavior, I consider the following specification, where I incorporate teacher and classroom fixed effects, as well as other observed characteristics:

$$\overline{\theta}^{T}_{jst} - \theta_{ijst} = g(\theta_{ijst}) + X'_{ijst}\gamma + \phi_j + \tau_{ct} + \epsilon_{ijst} \tag{2.2}$$

$g(\theta_{ijst})$ is a third-order polynomial in the corresponding test score in ninth-grade. In addition to gender and race, X_{ijst} includes a cubic polynomial of the seventh and eighth-grade test scores in math and English, the number of out-of-school suspensions, absences, and an indicator if the student was held back in seventh and eighth grades. Finally, I also include GPA in eighth grade. I control for unobserved teacher and classroom inputs by using teacher

(ϕ_j) and classroom fixed effects (τ_{ct}). Therefore, this specification compares differences in assessments across students of different gender and race within the same classroom after controlling for time-invariant characteristics of teachers.

Table 9 shows the estimates separately by subject. Columns (1), (4), and (7) show that, after accounting only for contemporaneous test scores and unobserved teacher and classroom characteristics, females are rated on average 0.14 s.d. higher in math, 0.05 s.d. higher in English, and 0.07 s.d. higher in Biology, relative to same-race boys. In contrast, black and Hispanic students are underrated in English and Biology. After accounting for lagged scores and behavioral controls, we observe a reduction in some of these estimates. Column (2) shows that the estimate for girls does not vary, but the differences across race become insignificant. Columns (5) and (8) show that the differences across race decrease by around half for English and Biology, but they are still statistically significant at the 1% level. Finally, the last set of estimates also includes indicators if the teacher and student exhibit a demographic match. Relative to the previous set of estimates, the inclusion of these additional controls does not change substantially the interpretation of the results. Overall, the data shows that teachers systematically overassess girls in the three subjects I consider. In terms of differences by race, I also find evidence for underassessment in English and Biology. In the case of math teachers, the unadjusted differences vanish once lagged scores and behaviors are accounted for.

Finally, Table A1 in the Appendix shows the estimates of a specification that also controls for the assessment of the teacher who taught the student in the previous year (8th grade), $T_{ijs,t-1}$. Although teachers are required to base their assessments solely on academic performance, as stated in Figure 1, the inclusion of this additional variable alleviates concerns related to the existence of unobserved characteristics, not captured by X_{ijst}, that teachers can consider to base their judgments and evaluate students. The results show small differences in the gender and ethnicity indicators, validating the patterns observed in Table 9.

Overall, these descriptive results are consistent with previous literature documenting similar patterns across different countries and education levels.[15] Based on this evidence, I proceed to study the impacts of this teacher capability, which is the main objective of this paper. The next section provides a simple framework to motivate the empirical strategy I employ in this paper.

3 Conceptual Framework

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:

$$\Delta\theta_i = \beta e_i + \alpha X_i + \phi_j^{VA} \tag{3.1}$$

Equation 3.1 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

[15] 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 underassessing increases by 2.5 and 3.5 percentage points for black Caribbean students in English and Science, respectively. Hanna and Linden (2012) show that teachers randomly assigned to grade exams rated low caste children in India between 0.03 and 0.08 standard deviations lower than high caste children. Cornwell et al. (2013) show that elementary teachers rate more favorably girls than boys, after accounting for test scores, but these differences are largely accounted for noncognitive skills. Rangel and Shi (2021) 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.

is a costly task in terms of effort, and I consider two sources of heterogeneity for this cost. First, evaluating 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 accurately. To incorporate this element, I assume that teacher beliefs about a student's skill depend partly on the student observed characteristics, X_i. Second, teachers differ in their ability to assess students. I model this heterogeneity using a fixed parameter ϕ_j^I, which captures each teacher's bias to evaluate students. This parameter shifts the cost function so that teachers with $\phi_j^I < 0$ will optimally choose to underrate all students, regardless of their characteristics. Imposing $\phi_j^I = 0$ and $\gamma = 0$ implies that all teachers are unbiased.[16] Taking into consideration these points, I assume that each teacher chooses an assessment for each student, based on the following minimization problem:

$$\min_{T_{ij}} \quad \frac{(T_{ij} - (\theta_{i0} + \gamma X_i))^2}{2} - T_{ij}\phi_j^I \tag{3.2}$$

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

$$T_{ij} = \theta_{i0} + \gamma X_i + \phi_j^I \tag{3.3}$$

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 $\pi_i = \pi(X_i) \in [0, 1]$:

$$\hat{\theta}_i = \pi_i \theta_{i0} + (1 - \pi_i) T_{ij} \tag{3.4}$$

After the student chooses the effort level $e(\hat{\theta}_i)$, the final skill level corresponds to $\theta_{i1} =$

[16] I model ϕ_j^{VA} as a common input that every student receives. Teachers could also impact students through heterogeneity in instruction, implying a correlation between ϕ_j^{VA} and ϕ_j^I. 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 academic teaching practices in the classroom, so I abstain from incorporating this channel into the model.

$\theta_{i0} + \Delta\theta_i$. Observed outcomes are a function of θ_{i1}:

$$y_i = \alpha^Y + \beta^Y \theta_{i1} + \epsilon_i \tag{3.5}$$

Substituting $e(\hat{\theta}_i)$ into (3.1) leads to a reduced-form equation which relates outcomes to the teacher characteristics, ϕ_j^{VA} and ϕ_j^I, as well as to the other inputs of the educational process:

$$y_{ij} = \delta_1 \theta_{i0} + \delta_2 X_i + \delta_3(X_i)\phi_j^I + \delta_4 \phi_j^{VA} + \epsilon_{ij} \tag{3.6}$$

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 $\delta_4 = \beta^Y$. This simple model highlights how assessments can influence skill accumulation and later outcomes.[17] From the teacher perspective, some students are more difficult to assess correctly than others. 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 (3.6) 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. This prediction motivates my empirical strategy to estimate teacher assessment skills in the data and measure their

[17] For expositional clarity, I have emphasized the role of assessments on skill beliefs. This model can be easily adjusted to incorporate other potential channels, such as differences in instruction, feedback, or motivation.

[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.

impact across different groups of students, which I discuss in the next section.

4 Empirical Analysis

My empirical strategy consists of two steps. First, I construct empirical Bayes estimates of assessment practices for each teacher, using each student's academic level as a reference point. The goal of this part is to isolate each teacher's persistent component from students' characteristics and other school and classroom-level characteristics. The second step consists of projecting these teacher-level estimates onto different outcomes and test whether these have heterogeneous impacts across gender and race-ethnicity subgroups.

4.1 Estimating Teacher Inaccuracy

Let $T_{ijst} \in \{1, 2, 3, 4\}$ be the assessment of student i reported by teacher j in the school-subject combination s, and year t, and $A_{ijst} \in \{1, 2, 3, 4\}$ be the observed achievement level of this student. Recall that A_{ijst} is a deterministic function of the test score θ_{ijst}, with each cutoff determined every year by the State Board of Education. T_{ijst} and A_{ijst} are discrete variables while θ_{ijst} 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 define an assessment *deviation* as the difference between the average score of all students rated by teacher j in level T_{ijst} in that year and student i's test score:[19]

$$D_{ijst} = \sum_{k \in \mathcal{K}_{(i)}} \frac{\theta_{kjst}}{N_{\mathcal{K}}} - \theta_{ijst} = \overline{\theta}^{T}_{jst} - \theta_{ijst}$$

Where $\mathcal{K}_{(i)} = \{k : T_{kjst} = T_{ijst}\}$ represents the set of students in the same school, year and subject who received the same assessment as i, and $N_{\mathcal{K}}$ denotes the number of students in this set. To illustrate this definition, consider an English teacher who rates a given student i in level 2. In this case, D_{ijst} will be the difference between the average score of students

[19] This definition also allows to use a different statistic, for example the median. In additional analyses (not reported) I experimented with the use of the median to define the deviation from the student test score. The main results do not change substantially.

in level $T_{ijst} = 2$ in that year and subject, and the actual score obtained by the student. Therefore, D_{ijst} captures the relative difference between the average test score associated to teacher j's assessment and student i's test score, which serves as a reference point. Employing D_{ijst} as a measure of deviation has two main advantages over the use of $T_{ijst} - A_{ijst}$. On the one hand, the distribution of D_{ijst} has a larger support than the (discrete) distribution of $T_{ijst} - A_{ijst}$. On the other hand, it is expressed in test score standard deviations, which facilitates its interpretation. Figure 7 displays the distribution of D_{ijst} separately by subject. Each subplot shows substantial variation across students. In additional analyses, I compare the discrete values T_{ijst} and A_{ijst} observed for each student-teacher pair to define the following binary variables:

Underassessment:

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

Overassessment:

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

Based on the definition of D_{ijst}, I construct empirical Bayes estimates of each teacher capability to assess students, accounting for non-random student assignment into classrooms. With this objective in mind, I start by estimating the following regression, separately by subject:

$$D_{ijst} = X'_{ist}\gamma + C'_{ijst}\delta + \phi^I_j + \tau_s + \epsilon_{ijst} \tag{4.1}$$

Equation (4.1) applies a teacher value-added specification to isolate a teacher-specific component determining variation in D_{ijst}, net of student and school characteristics.[20] Following

[20] Hill and Jones (2021) employ a similar approach to recover a teacher-specific measure of optimism, by using T_{ijst} as the dependent variable and including student-subject fixed effects. My approach is different in two aspects. First, they employ these teacher fixed effects as an instrument to estimate the impact of a higher value of T_{ijst} on student i's contemporaneous test scores. I do not attempt to estimate the direct effect of increasing an assessment on the same student's outcomes. Instead, my goal is to estimate the effect of exposure to teachers whose judgments are more likely to differ from the scholastic aptitudes captured by test scores. My empirical strategy to answer this question relies on a different set of assumptions. Second, since I focus on ninth-grade students, equation (4.1) uses only teacher fixed effects, exploiting variation across different cohorts to identify ϕ_j. This approach alleviates concerns related to the identification of teacher effects under dynamic sorting.

this literature (Kane and Staiger, 2008; Chetty et al., 2014; Jackson, 2018; Petek and Pope, 2021), this specification includes as student background characteristics X_{ist} indicators for gender, race, and parental education, a cubic polynomial of the seventh and eighth-grade test scores in math and language, number of days suspended out of school in seventh and eighth grades, absences in seventh and eighth grades, and GPA in eighth grade. C_{ijst} are leave-one-out, classroom-level, average characteristics of student i's peers (share of peers by race and gender; share of peers by parental educational level; average scores in math and reading in 8th grade; average number of suspensions in 8th grade; share of repeating students; share of economically disadvantaged students). τ_s and ϕ_j^I correspond to a full set of school and teacher fixed effects. Therefore, this specification partials out any teacher time-invariant determinant of D_{ijst} to identify the parameters γ and δ.

After estimating $\hat{\gamma}$, $\hat{\delta}$, and $\hat{\tau}_s$ using OLS, I construct residuals $\varepsilon_{ijst} = D_{ijst} - X'_{ist}\hat{\gamma} - C'_{ijst}\hat{\delta} - \hat{\tau}_s$. Following Kane and Staiger (2008) and Jackson (2018), I assume that this residual can be decomposed into a teacher-specific component (ϕ_j), a classroom-specific shock (ε_{ijstc}), and a student-specific shock (ξ_{ijst}), so that $\varepsilon_{ijst} = \phi_j + \varepsilon_{ijstc} + \xi_{ijst}$. Under a selection-on-observables assumption, the average of residuals at the teacher-level, $\bar{\varepsilon}_j$, is an unbiased estimate of teacher j's contribution to the outcome D_{ijst}. To avoid mechanical endogeneity, I construct leave-year-out average residuals and compute empirical Bayes estimates of teacher effects in each year $\hat{\phi}_{jt}$ by using the leave-year-out average residuals $\bar{\varepsilon}_{j,-t}$ weighted by an estimate of reliability:[21]

$$\hat{\phi}_{jt} = \bar{\varepsilon}_{j,-t} \times \frac{\hat{\sigma}^2_{\phi}}{\hat{\sigma}^2_{\phi} + \hat{\sigma}^2_j} \tag{4.2}$$

Where:

$$\hat{\sigma}^2_j = \left[\sum_{m_j} \left(\hat{\sigma}^2_{\varepsilon_{ijstc}} + \frac{\hat{\sigma}^2_{\xi_{ijst}}}{N_{cj}} \right)^{-1} \right]^{-1} \tag{4.3}$$

I use a similar approach to recover estimates of teacher test score value-added, which I label

[21] This weighting parameter corresponds to the ratio between the estimate of the variance of ϕ_j ($\hat{\sigma}^2_{\phi}$) across all teachers and the estimated variance of the error ($\hat{\sigma}^2_j$). See the Appendix A.1 for details about the computation of each variance term.

$\hat{\phi}_{jt}^{VA}$. To compute these objects, I use (4.1) replacing the end-of-course test score of each subject as the dependent variable. This set of additional estimates allows me to control for teacher quality in the second part of the estimation. I present the distribution of the shrunken estimates of inaccuracy ($\hat{\phi}_{jt}^{I}$) and test score value-added ($\hat{\phi}_{jt}^{VA}$) in Figure 8, and the estimates of underassessment and overassessment in Figure A1. Table 7 presents the correlation between the different empirical Bayes estimates. Test score value-added is weakly correlated to both inaccuracy and precision, suggesting that the ability to raise test scores does not capture the ability to predict students' achievement. It also suggests that students exposed to low-value added teachers do not mechanically associate with exposure to less accurate teachers. Moreover, the correlation between inaccuracy and precision is -0.21, suggesting that these two measures reflect separate dimensions of teachers' ability to predict their students' achievement.

Table 8 shows the standard deviation of the distribution of each set of empirical Bayes estimates, separately by subject.[22] The first column shows the estimates for test score value-added. An increase of one standard deviation corresponds to an increase of between 0.014 and 0.049 test score s.d., depending on the subject. These values are in line with previous findings of teacher effectiveness in high school.[23] The second column shows that an increase of one standard deviation in the teacher inaccuracy distribution ($\hat{\phi}_{j}^{I}$) corresponds to an increase of 0.09 test score s.d. in the assessment deviation for English and math teachers. The distribution for Biology teachers exhibits a slightly lower standard deviation of 0.08 test score s.d. The third and fourth columns show the estimated standard deviation of the distribution of the empirical Bayes estimates for underassessment and overassessment. The standard deviation for underassessment ranges between 0.042 and 0.049, implying that an increase of one s.d. in the value of ϕ_{j}^{U} corresponds to an increase of 4.2 p.p. and 4.9 p.p. in the probability of a teacher underassessing all students on average. In the case of overassessment, the standard deviation is 0.023 and 0.037 for biology and math teachers, respectively, implying an increase of 2.3 p.p. and 3.7 p.p. in the probability of overassessment for all

[22] I describe the procedure to compute the variance of each teacher effect in Appendix A.1.

[23] In math, the estimated effect of increasing one s.d. in test score value-added ranges from about 0.08 to 0.21, and in English it ranges from 0.03 to 0.10 (Aaronson et al., 2007; Jackson, 2014; Mansfield, 2015)

students.

4.2 Identification

The primary identification challenge in recovering estimates of teacher effects in equation (4.1) 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 effects. 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^I 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, unobserved characteristics of teachers and students are uncorrelated with assignment. Therefore, I make the following conditional independence assumption:

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

Under this assumption, conditional on the set of controls, teacher j's capabilities are uninformative about the expected characteristics of students taught by this teacher. Thus, the conditional difference in the outcomes between teacher j and j' will yield the difference in the persistent inaccuracy and precision between teacher j and j'. I present evidence to support the validity of this assumption in section 6.1. 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.[24] Furthermore, the classroom-level observed characteristics C_{ijst} account for sorting at the group level based on similar characteristics.

[24] Previous studies typically consider two lags of test scores to account for selection based on ability (Rothstein, 2010; Jackson, 2014).

4.3 Estimation of Impacts on Student Outcomes

After constructing empirical Bayes estimates for each teacher, I estimate the following regression, which relates students' outcomes with the leave-year-out estimates of teacher inaccuracy, conditioning on teacher test-score value added and other covariates:

$$y_{ijst} = \beta_0 \hat{\phi}^I_{j,-t} + \beta_1 \hat{\phi}^I_{j,-t} \times \text{Girl}_i + \beta_{2e} \hat{\phi}^I_{j,-t} \times \text{Ethnic}_i + \gamma \hat{\phi}^{VA}_{j,-t} + X'_{ijst}\delta + \tau_s + \tau_t + \epsilon_{ijst} \quad (4.4)$$

In (4.4), the parameters of interest are represented by the vector $\beta = \{\beta_0, \beta_1, \beta_2\}$, which indicates the relationship between each outcome and changes in teacher inaccuracy across different subgroups of students. Each leave-year-out empirical Bayes estimate $\{\hat{\phi}^I_j, \hat{\phi}^{VA}_j\}$ is normalized so that β and γ can be interpreted as the effect of increasing each teacher estimate by one standard deviation. The term Girl_i is a binary variable equal to one if the student is female and Ethnic_i is a vector of indicator variables for a student's ethnicity. The categories for ethnicity are the following: black, Hispanic, Asian, and other (Pacific Islander, Indian, and Multiple races). Therefore, the estimate β_1 indicates the differential impact for girls relative to boys, while β_{2e} indicates the differential impact for a given ethnicity relative to white students. X_{ijst} corresponds to the same student-level and classroom-level controls employed in the initial step. I use my estimates of test-score value added $\hat{\phi}^{VA}_{j,-t}$ to control for teacher quality. Additionally, to account for tracking (Jackson, 2014), I include a set of school-track fixed effects τ_s, which allow for comparisons within a set of students in the same school taking the same classes in ninth-grade. I define a track as the combination of the following courses: English I, Algebra I, Introduction to Algebra, Geometry, Biology, Earth/Environmental Sciences, World History, and Spanish I.[25] Finally, I cluster standard errors at the teacher and student levels to account for cases where a student is linked to more than one teacher.

[25] Additionally, for English I and Algebra I, I classify each class by academic level (regular, basic, honors).

I analyze the following 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.

5 Results

This section presents the main findings of the paper. In the first subsection, I analyze the heterogeneous impacts of exposure to teachers with different propensity to assess students on contemporaneous outcomes. Then, I present my results for outcomes observed at the end of high school.

5.1 Impact on Contemporaneous Outcomes

I start my analysis by presenting the estimates of $\{\beta_0, \beta_1, \beta_{2e}\}$ in equation (4.4) on contemporaneous outcomes observed at the end of ninth grade. Specifically, I consider end-of-grade test scores in the same subject and an indicator equal to one if the student expects to attend a two-year or a four-year college after graduating. Table 10 shows the estimates of (4.4) for $\hat{\phi}_j^I$ and the interaction terms with the gender and race-ethnicity indicators. Each column also includes the estimate of test score value-added, $\hat{\phi}_j^{VA}$. Columns (1) and (2) in Table 10 show that an increase of 1 s.d. in the distribution of $\hat{\phi}_j^I$ (that is, exposure to a more inaccurate teacher) impacts girls positively, relative to boys. The interaction coefficient is 0.004 (p-value<0.05). This estimate is equivalent to an increase of 0.1 s.d. in the teacher value-added distribution. Column (2) analyzes differential impacts across ethnicity groups and shows that only the interaction term for other-race students is statistically significant at the 5% level. Figures 9 and 10 display graphically the estimates of the total marginal effect and their 95% confidence interval for each subgroup. In the case of contemporaneous test scores, the left-hand side plot in each figure shows that girls, as well as black and Asian students, are positively affected by increases in $\hat{\phi}^I_j$. However, the estimates across ethnic groups are less precise. Column (3) shows that more inaccurate teachers increase girls' college aspirations and decrease boys'. An increase of 1 s.d. in the distribution of $\hat{\phi}^I_j$ implies a reduction of 0.3 p.p. (p-value<0.01) in the probability of planning to attend college for boys and an increase of 0.2 p.p. (p-value<0.01) for girls. Similar to contemporaneous test scores, I do not find large differences when examining heterogeneous effects by ethnicity. Column (4) shows that the main effect is only statistically significant at the 10% level, and the only interaction term that is statistically significant at the 5% level is other-race. The right-hand side plots in Figures 9 and 10 show that the marginal effects across subgroups are less precise and not statistically different from zero at the 10% level.

5.2 Impact on 12th Grade Outcomes

Table 11 shows the results of the estimation of (4.4) for outcomes observed three years after exposure. Specifically, I present the estimates for the (weighted) 12th grade GPA score reported by each school, the intention to attend a two-year or four-year college, an indicator equal to one if the student took the SAT, and the SAT score, conditional on taking the test. Similarly to the previous tables, in each column I report the estimates of $\hat{\phi}^I_j$ as well as the estimate of its interaction with the student's gender or race-ethnicity indicator, also controlling for teacher value-added ($\hat{\phi}^{VA}_j$). I also present graphically the marginal effects separately for each subgroup in Figures 11 and 12. Column (1) in Table 11 shows that, conditional on teacher value-added, an increase of 1 s.d. in the distribution of $\hat{\phi}^I_j$ has a differential impact of 0.003 (p-value<0.1) for girls relative to boys. This estimate implies a marginal increase of 0.003 (p-value<0.05) points for girls and no impact for boys. The differences across ethnoracial categories are significantly larger. The interaction terms in column (2) show that, relative to white students, exposure to a teacher who is more likely to overassess students positively impacts black, Hispanic, and Asian students. Columns (3) and (4) show the estimates of expectations about college attendance reported in 12th grade. As in my results of 9th grade expectations, I find evidence of heterogeneous impacts across gender but not substantial differences by ethnoracial groups. An increase of 1 s.d. in $\hat{\phi}_j^I$ decreases by 0.2 p.p. the probability of expecting to attend college for boys, but it increases this probability by the same amount for girls. The corresponding plot in Figure 11 shows that the marginal effects for both groups are statistically significant at the 1% level. As in the case of college expectations reported in 9th grade, I find small differences across race and ethnicity. The interaction terms in column (4), as well as the plot in Figure 12, show a marginal effect of 0.5 p.p. for Asian students (p-value<0.05), while the other subgroups have much smaller and less precise estimates. Columns (5)-(8) analyze the effects on whether a student took the SAT as well as her total score. Regarding SAT taking, column (5) shows that an increase of 1 s.d. in the distribution of $\hat{\phi}_j^I$ induces a positive increase for girls in the probability of taking the SAT. Still, the total marginal effect is not statistically different from zero. On the contrary, I find differences between white and non-white students. The marginal effect for black and Hispanic students is around 0.15 p.p. (p-value<0.1), while the marginal impact for Asian students is 0.64 p.p. (p-value<0.05). Finally, columns (7) and (8) in Table 11 show that, conditional on taking the exam, exposure to more inaccurate teachers has a positive impact for girls, black, and Hispanic students but a negative impact for Asian students. Interestingly, the estimates from column (8) show a negative effect for Asians of around 3.5 SAT points (p-value<0.01). This pattern suggests that while exposure to more inaccurate teachers positively impacts school-level achievement (GPA) and a measure of human capital accumulation (SAT) for black and Hispanics, only the former is true for Asian students. This last pattern is consistent with psychology literature studying the different impacts of positive stereotypes (Czopp et al., 2015). In particular, positive stereotypes about Asian students' skills may lead to hinder performance when high expectations are made salient by an outgroup member (Cheryan and Bodenhausen, 2000).

One aspect worth mentioning about these results is that the measure of teacher inaccuracy is more predictive for 12th grade outcomes than teacher 9th grade test score value-added. This point is consistent with recent evidence showing the importance of non-test score teacher quality dimensions. For example, Jackson (2018) finds that 9th grade behaviors teacher value-added or 10th grade GPA value-added increase the variance explained by teacher effects by a large fraction.[26] To quantify the increase in explanatory power, I conduct a similar exercise by computing the change in explained variance between the baseline scenario (where only $\hat{\phi}_j^{VA}$ is included in (4.4)) compared to the case where $\hat{\phi}_j^{I}$ is included as an additional regressor.[27] Table 12 shows the change in explained variance for several outcomes. Except for contemporaneous test scores, there are substantial increases in almost every outcome. In particular, those associated with motivation or expectations, such as the intention to attend college or 12th grade GPA.

In addition, it is important to consider the magnitude of these estimates. Although small, relative to the average values of each outcome, the estimates are similar to the effects of increases in non-test score dimensions of teacher effectiveness, such as behavior or learning skills teacher value-added. For example, Jackson (2018) finds that an increase of 1 s.d. in behavior teacher value-added in 9th grade increases GPA in twelfth grade by 0.021 points. His estimates for SAT taking and SAT scores are 0.012 and -0.232, respectively (Table 7, page 2102). Using data from Los Angeles school districts, Petek and Pope (2021) 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, SAT scores by 2 points and GPA at the end of high-school by 0.013 points (Table 5, page 56). Moreover, since these estimates are based on exposure in a specific grade, they potentially underestimate the total effect if we consider, for example, exposure to a teacher with a similar propensity to mis-assess specific groups for a longer number of years.

Taken together, these results show that the descriptive findings discussed in section 2.3 have implications for outcomes observed three years later. While the differences across gender are consistent with the existing literature about gender-specific impacts of teachers who are more positively biased towards one group (Lavy and Sand, 2018; Lavy and Megalokonomou,

[26] For example, Jackson (2018) finds increases up to 793% in explained variance and up to 151% when behavior value-added and 10th Grade GPA value-added are included in a model only considering test score value-added, respectively.

[27] More specifically, I calculate $\left(\frac{Var(\hat{\gamma}_1\hat{\phi}_j^{VA}+\hat{\gamma}_2\hat{\phi}_j^{I})}{Var(\hat{\gamma}_1\hat{\phi}_j^{VA})}-1\right)$ and present this ratio in percentage points for different outcomes.

2019; Terrier, 2020), my findings also emphasize the consequences of these mis-assessments across racial and ethnic groups, particularly for black, Hispanic, and Asian students.

6 Robustness Checks

In this section, I check the validity of the main results presented in the previous section by conducting three different robustness checks. First, in subsection 6.1, I test the robustness of my strategy to student sorting. Then, in subsection 6.2, I account for the potential influence of other teachers by using a sub-sample of students linked to a Math and English teacher and employing an additional set of fixed effects. Finally, in subsection 6.3, I check whether the main patterns are robust to the use of alternative measures of mis-assessment.

6.1 Testing for Student Sorting

One natural concern is that principals can assign teachers to different classrooms based on unobserved characteristics, violating the conditional randomness assumption. If this is true, then the estimates discussed in the previous section could merely reflect student sorting. One alternative to address this possibility is to test whether, conditional on the main controls, students with higher predicted outcomes (based on background characteristics and achievement in 7th grade) are systematically assigned to different types of teachers in 9th grade. To test this possibility, I regress each outcome onto background characteristics (gender, ethnicity, parental education), achievement in 7th grade (a third-degree polynomial on ELA and math scores), and behaviors in 7th grade (absences, suspensions, and being held back) to create a predicted outcome. Then, I regress each predicted outcome onto 8th-grade controls and the set of fixed effects. Table 13 shows that there is no evidence of sorting of students to teachers with higher or lower accuracy.

Nevertheless, there is still the possibility of selection based on unobserved characteristics. For instance, students with particular characteristics not captured by lagged 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 employ an instrumental variables (IV) strategy. I follow Rivkin et al. (2005) and aggregate the level of treatment to the school-year level and use this variable as an instrument for each teacher fixed effect. This strategy exploits within-school, across-cohort variation in the composition of teachers to identify the effects of exposure. I compare students belonging to a cohort where, on average, teachers are more likely to deviate from what test scores indicate relative to other cohorts with a different teacher composition in the same school. Since the selection of students across cohorts based on these teacher characteristics is unlikely, aggregating the treatment at the school-year level helps overcome the selection-on-unobservables problem. If the estimates reported using the specification 4.4 are a consequence of sorting based on unobserved characteristics, then we should expect the estimates obtained using the IV strategy to be much smaller.

Tables 14 and 15 show the IV estimates. The size and sign of the estimates of inaccuracy and the interaction terms are similar to the main analysis, particularly for 12th grade outcomes. Therefore, this analysis suggests that the main results reflect exposure to teachers who differ in how they assess students, relative to test scores, and are not a consequence of selection of students to teachers, based on either observed or unobserved characteristics.

6.2 Accounting for Other-Subject Teachers

In addition to concerns related to non-random selection, it could be possible that some of the effects captured by the estimates in the main specification reflect the exposure to other teachers in different subjects. To check the robustness of my results to this alternative explanation, I re-estimate equation (4.4) also including a set of other-subject teacher fixed effects for the sub-sample of students linked to more than one teacher. This sub-sample consists of approximately 60% of the total number of observations. Tables 16 and 17 show the estimates for 9th grade and 12th grade outcomes, respectively. Overall, the patterns are qualitatively similar to the main results presented in Tables 10 and 11. This test suggests that the main specification adequately captures teachers' individual impact.

6.3 Using Alternative Measures of Teacher Assessments

In this section, I consider two alternative measures to characterize how teachers persistently assess students over time and whether the main patterns discussed in the previous section hold. As mentioned in section 4, instead of relying on a continuous measure it is possible to compare the discrete values T_{ijst} and A_{ijst} observed for each student-teacher pair and define the following binary variables:

Underassessment:

$$\mathbb{1}\{T_{ijst} < A_{ijst}\} \tag{6.1}$$

Overassessment:

$$\mathbb{1}\{T_{ijst} > A_{ijst}\} \tag{6.2}$$

I employ a similar specification to (4.1) to estimate each teacher's empirical Bayes estimates. Using (6.1) and (6.2) as dependent variables implies that the teacher fixed effect will now capture each teacher's propensity to judge students' mastery of the subject below or above their achievement level, regardless of the magnitude of this difference. Moreover, defining these two variables separately allows to test whether their effects are symmetric. Let $\hat{\phi}_j^U$ and $\hat{\phi}_j^O$ denote each teacher j's empirical Bayes estimate of underassessment and overassessment, respectively. Table A2 shows a selection-on-observables test similar to the one discussed in section 6.1 for the main inaccuracy measure.

Figures A2-A3 in the Appendix display the results using the measure of underassessment to classify teachers, while Figures A6-A7 show the results employing the measure of overassessment. The main patterns discussed in section 5.1 remain consistent after conducting these additional analyses, although the coefficients are less precise. Overall, these additional analyses show that girls are positively impacted by exposure to teachers more likely to overassess (or less likely to underassess) students. In addition, non-white students also benefit from this behavior. Nevertheless, for Asian students, while we observe positive impacts on college expectations and grades, the effects on SAT scores are negative.

7 Conclusion

In this paper, I use North Carolina public schools data to study the relationship between teacher assessments received by ninth-grade students and several outcomes during high school. I employ information from the state-level standardized test scores and teacher assessments, collected when students are taking the standardized tests, to estimate teachers' persistent tendency to assess students above or below the academic level proxied by test scores. I employ these teacher-specific measures to estimate the impact of exposure to this dimension of teacher effectiveness across various outcomes observed in ninth and twelfth grades. This paper contributes to the literature about the impacts of teacher subjective evaluations by estimating the differential impacts of exposure to teachers who systematically tend to overrate or underrate students. It also connects to a broader literature that studies the multidimensionality of teacher quality.

I find that, conditional on test score value-added, teachers who are more likely to overrate students have positive effects for girls and black, Hispanic, and Asian students on their intention to attend college, GPA scores, and SAT taking. Nevertheless, these positive impacts do not translate to improvements in SAT scores for all these groups. I find positive and statistically significant estimates only for black and Hispanic students but negative and statistically significant estimates for Asian students. Moreover, these estimates are comparable to the recent literature documenting the effects of increasing non-test score dimensions of teacher value-added, such as behavior or learning skills.

This work leaves several avenues for future research. First, a natural question is related to how this teacher capability also impacts college attendance and major choice. Recent work shows that teacher gender biases in high school predict performance in university admission exams and selection of fields of study (Lavy and Megalokonomou, 2019). Analyzing whether mis-assessments also affect post-school choices dissimilarly across gender and race is a fruitful topic for further research. Second, a critical unexplored channel is the role of parents. Unfortunately, the North Carolina data does not include parental beliefs or measures of parental investments. Analyzing how parents react to the type of signals studied in this paper can increase our understanding about how parents influence children's effort and achievement.

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.

O. Attanasio, F. Cunha, and P. Jervis. “Subjective Parental Beliefs. Their Measurement and Role”. Working Paper 26516, 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*, 129(1):286–322, 2021.

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”. *The Quarterly Journal of Economics*, 134(3):1163–1224, 2019.

S. Cheryan and G. Bodenhausen. “When Positive Stereotypes Threaten Intellectual Performance: The Psychological Hazards of “Model Minority” Status”. *Psychological Science*, 11(5):399–402, 2000.

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.

A. Czopp, A. Kay, and S. Cheryan. "Positive Stereotypes Are Pervasive and Powerful". *Persepectives on Psychological Science*, 10(4):451–463, 2015.

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

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.

A. Hill and D. B. Jones. "Self-Fulfilling Prophecies in the Classroom". *Journal of Human Capital*, 15(3):400–431, 2021.

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.

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*, 129(1):81–100, 2021.

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

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.

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.

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.

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 Students”. Working paper, Department of Economics, University of Maryland, 2021.

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, 2021.

S. Rivkin, E. Hanushek, and J. Kain. “Teachers, Schools, and Academic Achievement”. *Econometrica*, 73(2):417–458, 2005.

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

Y. Shi and M. Zhu. “Model Minorities in the Classroom? Positive Bias Towards Asian Students and its Consequences”. Working paper, 2021.

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.

8 Figures and Tables

Table 1: Availability of Teacher Assessments

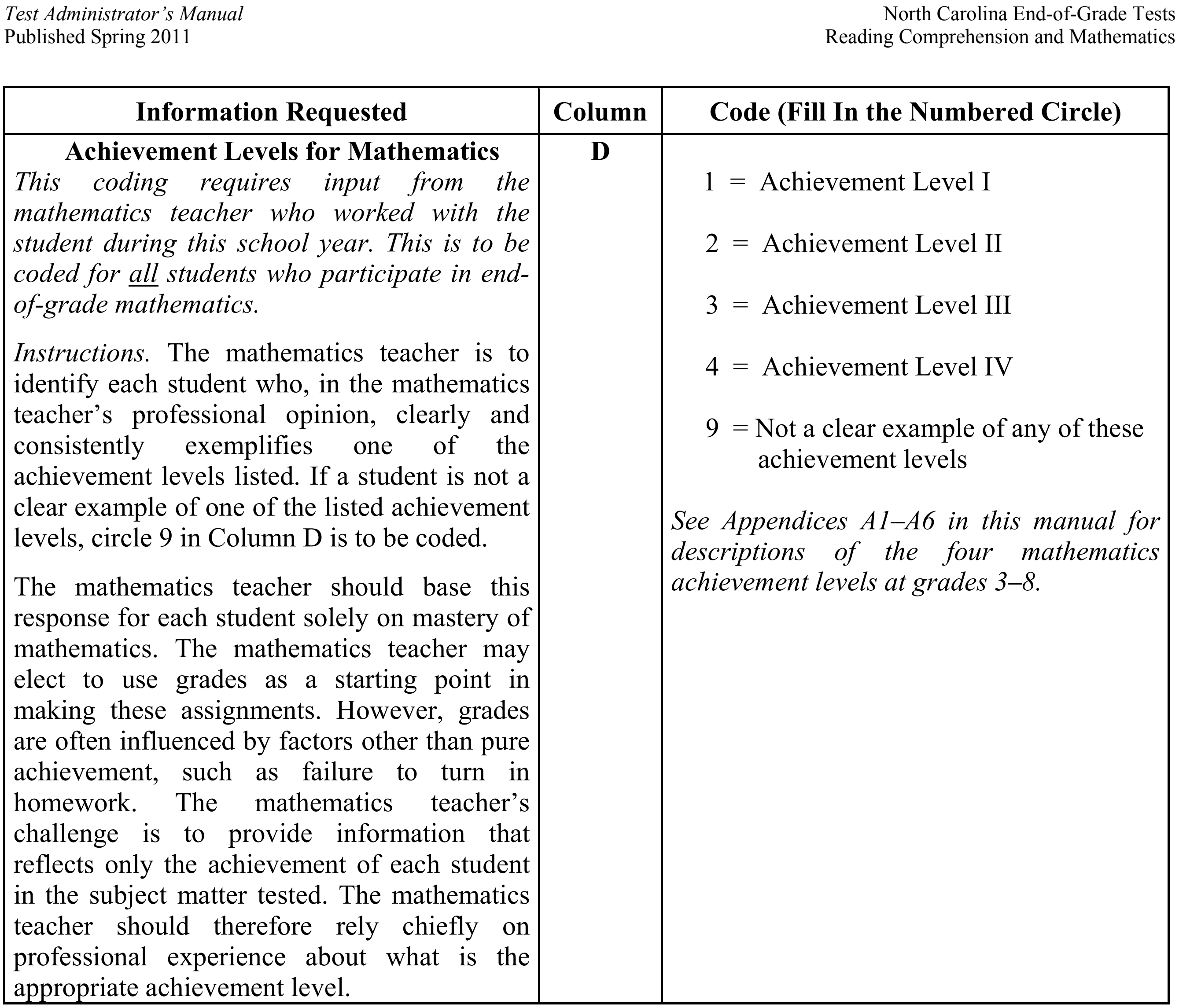
	2007	2008	2009	2010	2011	2012	2013
Algebra I	✓	✓	✓			✓	✓
Algebra II	✓	✓	✓	✓			
English I	✓	✓	✓	✓	✓	✓	
Biology			✓	✓	✓	✓	✓

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

	Raw Test Scores				
	2007-2012			2013	
	Algebra I	English I	Biology	Algebra I	Biology
Level IV: Superior	158-181	157-176	159-179	264-281	261-275
Level III: Consistent	148-157	146-156	147-158	253-263	252-260
Level II: Inconsistent	140-147	138-145	138-146	247-252	243-251
Level I: Insufficient	118-139	119-137	121-137	226-246	225-242

Notes: This table shows the range of raw test scores considered within each achievement level in the corresponding year. The third column considers end-of-course test scores for Biology during years 2009-2012.

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



Test Administrator's Manual
Published Spring 2011

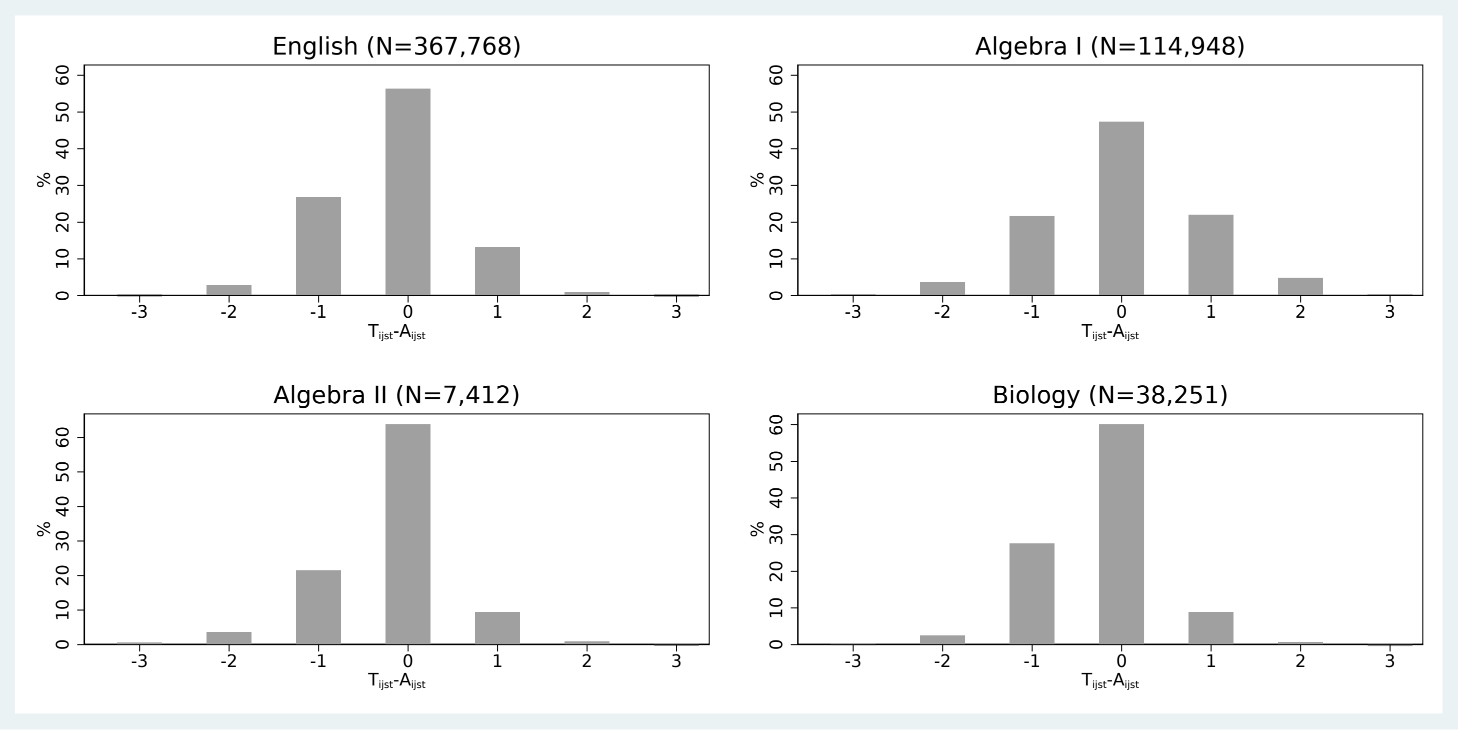
North Carolina End-of-Grade Tests
Reading Comprehension and Mathematics

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.*

Table 3: Summary Statistics.

Variable	Mean	Std. Dev.	Number Obs.
Unit of observation: Student			
White	0.58	0.49	476125
Black	0.27	0.44	476125
Hispanic	0.08	0.28	476125
Asian	0.02	0.15	476125
Algebra I Score	0.19	0.96	274490
English I Score	0.18	0.92	409593
Biology I Score	0.15	0.93	48731
English score (8th grade)	0.07	0.96	476125
Math score (8th grade)	0.08	0.96	476125
Repeated (8th grade)	0.01	0.08	472863
Days suspended out of school (8th grade)	0.24	1.61	476125
Repeated (7th grade)	0.01	0.09	476125
Days suspended out of school (7th grade)	0.17	1.42	476125
Days absent (7th grade)	6.68	6.77	476125
Times tardy (7th grade)	0.96	4.96	476125
$\mathbb{1}(T_{ijst} = A_{ijst})$ (Algebra I)	0.47	0.50	128113
$\mathbb{1}(T_{ijst} = A_{ijst})$ (English I)	0.56	0.50	383488
$\mathbb{1}(T_{ijst} = A_{ijst})$ (Biology I)	0.59	0.49	45724
Unit of observation: Teacher			
White teacher	0.84	0.36	5612
Black teacher	0.13	0.34	5612
Hispanic teacher	0.01	0.08	5612
Female teacher	0.76	0.43	5612
Avg. experience (years)	9.47	9.57	5607
Initial experience (years)	8.18	9.32	5605
Education: Bachelor's degree	0.71	0.45	5607
Education: Master's degree	0.28	0.45	5607

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 4: Conditional Distribution of A_{ijst}: English

	Achievement Level (A_{ijst})			
	Level I	Level II	Level III	Level IV
Teacher Assessment (T_{ijst})				
Level I	37%	14%	3%	0%
Level II	43%	42%	19%	4%
Level III	19%	42%	64%	44%
Level IV	1%	2%	14%	52%
	100%	100%	100%	100%

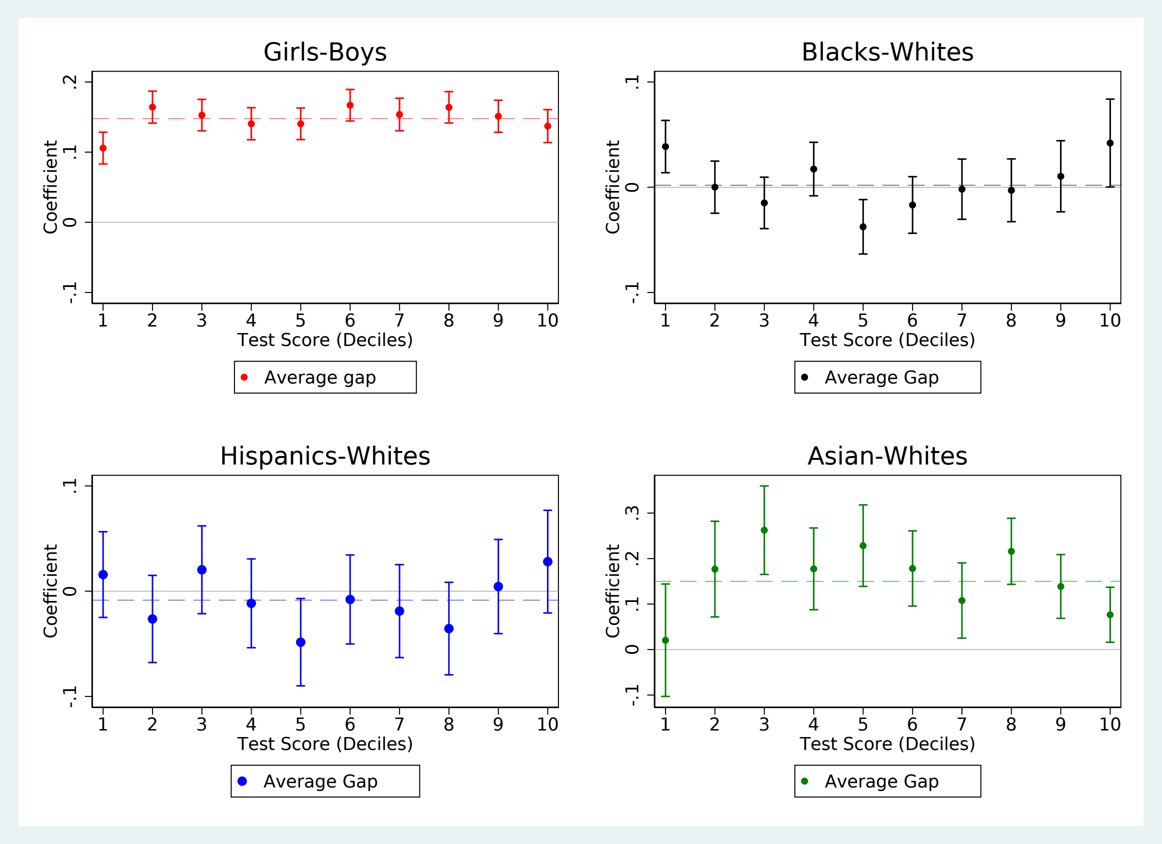
Table 5: Conditional Distribution of A_{ijst}: Algebra I

	Achievement Level (A_{ijst})			
	Level I	Level II	Level III	Level IV
Teacher Assessment (T_{ijst})				
Level I	35%	16%	5%	1%
Level II	38%	37%	21%	7%
Level III	25%	42%	56%	42%
Level IV	2%	6%	18%	51%
	100%	100%	100%	100%

Table 6: Conditional Distribution of A_{ijst}: Biology

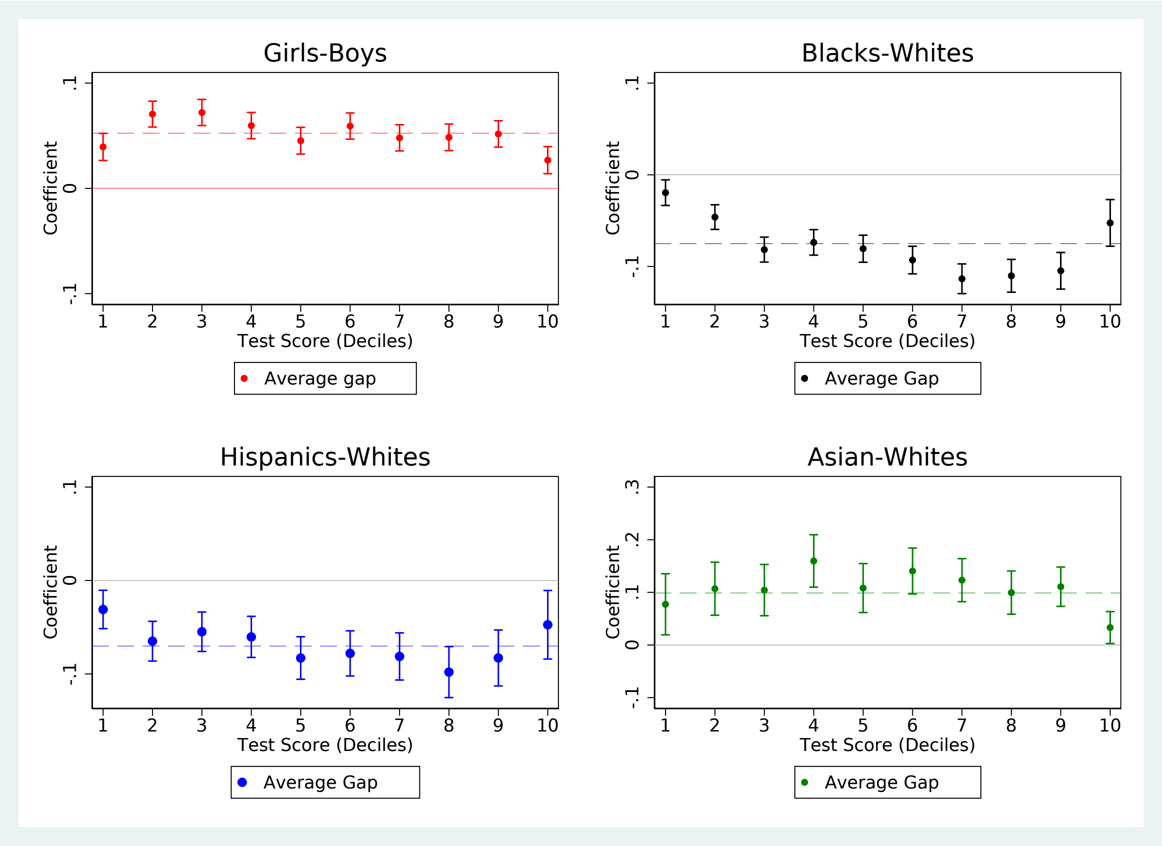
	Achievement Level (A_{ijst})			
	Level I	Level II	Level III	Level IV
Teacher Assessment (T_{ijst})				
Level I	49%	22%	3%	0%
Level II	37%	38%	18%	3%
Level III	14%	36%	62%	36%
Level IV	0%	4%	17%	61%
	100%	100%	100%	100%

Figure 3: Unadjusted Differences in Teacher Assessments: Math



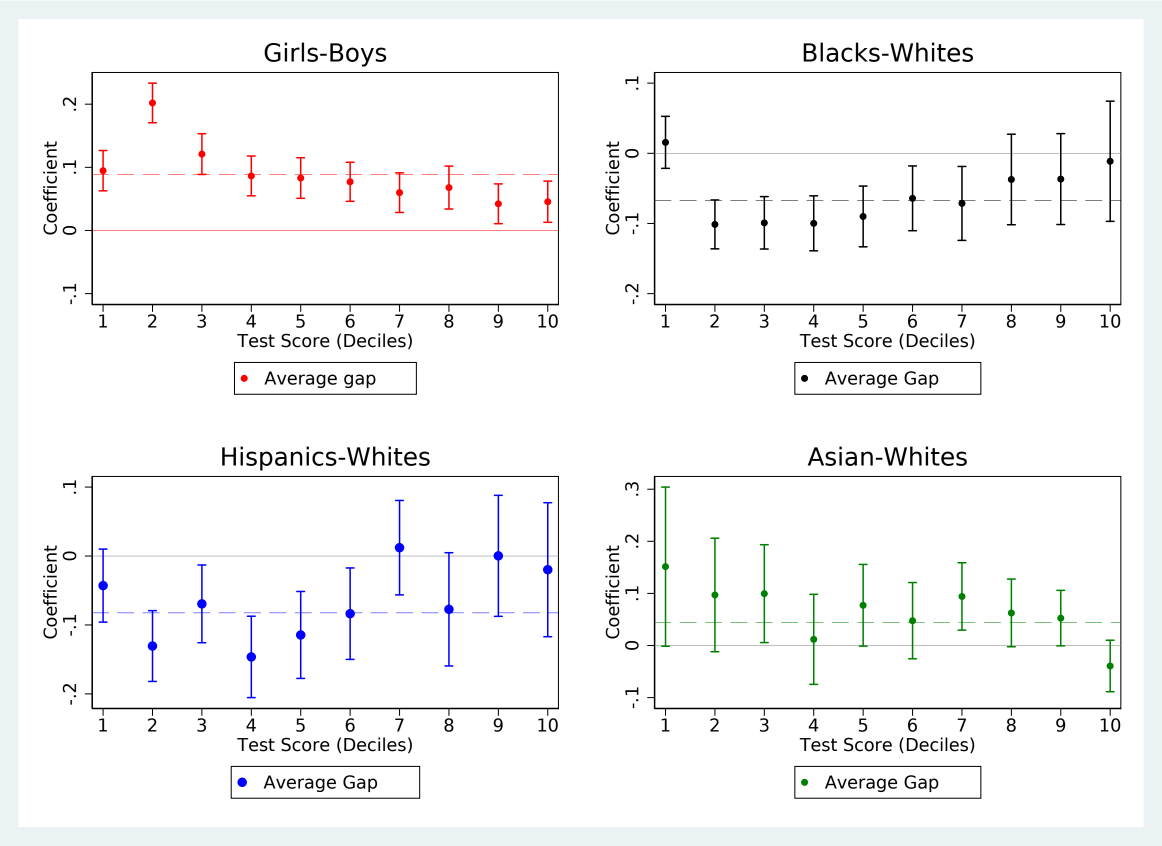
Notes: Each subplot shows the unadjusted differences in assessments between the corresponding groups (measured as test score standard devations) for each decile of the standardized math test score distribution. Each estimate corresponds to the coefficient $\beta + \gamma^d$ in (2.1). This estimation considers the total number of assessments for math courses available in the sample between 2007 and 2013.

Figure 4: Unadjusted Differences in Teacher Assessments: English



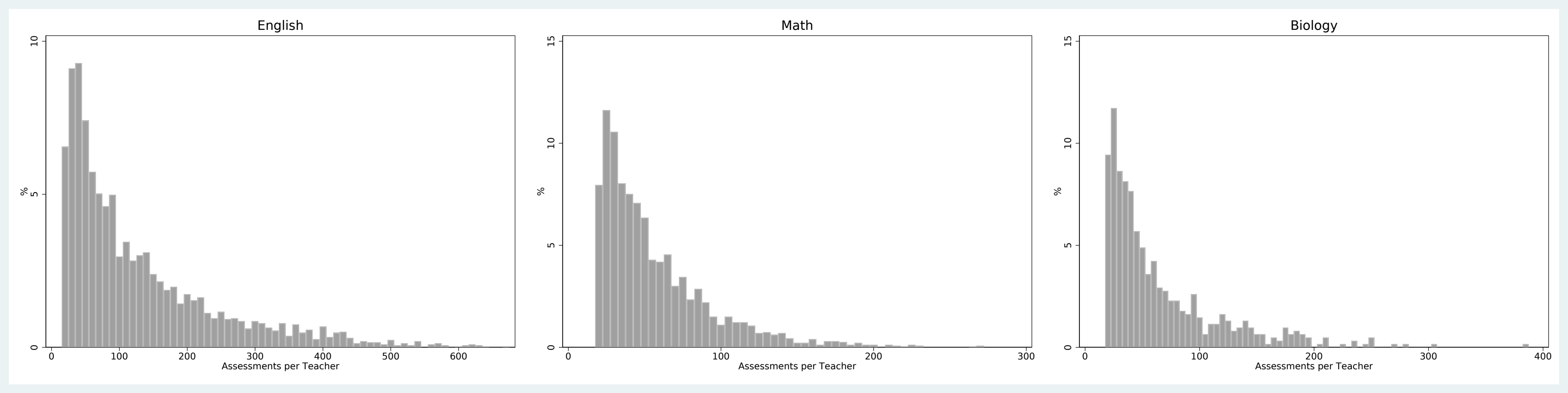
Notes: Each subplot shows the unadjusted differences in assessments between the corresponding groups (measured as test score standard devations) for each decile of the standardized English test score distribution. Each estimate corresponds to the coefficient $\beta + \gamma^d$ in (2.1). This estimation considers the total number of assessments for English I available in the sample between 2007 and 2012.

Figure 5: Unadjusted Differences in Teacher Assessments: Biology



Notes: Each subplot shows the unadjusted differences in assessments between the corresponding groups (measured as test score standard devations) for each decile of the standardized Biology test score distribution. Each estimate corresponds to the coefficient $\beta + \gamma^d$ in (2.1). This estimation considers the total number of assessments for Biology available in the sample between 2007 and 2013.

Figure 6: Number of Assessments Reported by Each Teacher



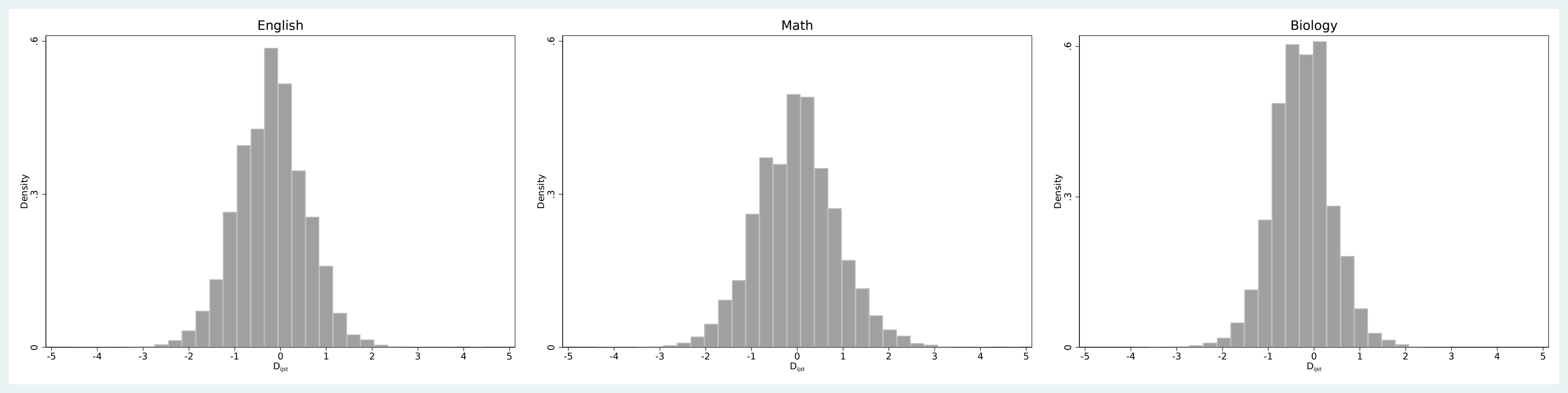
(a) English

(b) Math

(c) Biology

Notes: This plot shows the distribution of the number of assessments observed for each teacher for all years, after discarding teachers linked to less than 20 students.

Figure 7: Distribution of D_{ijst} by Subject



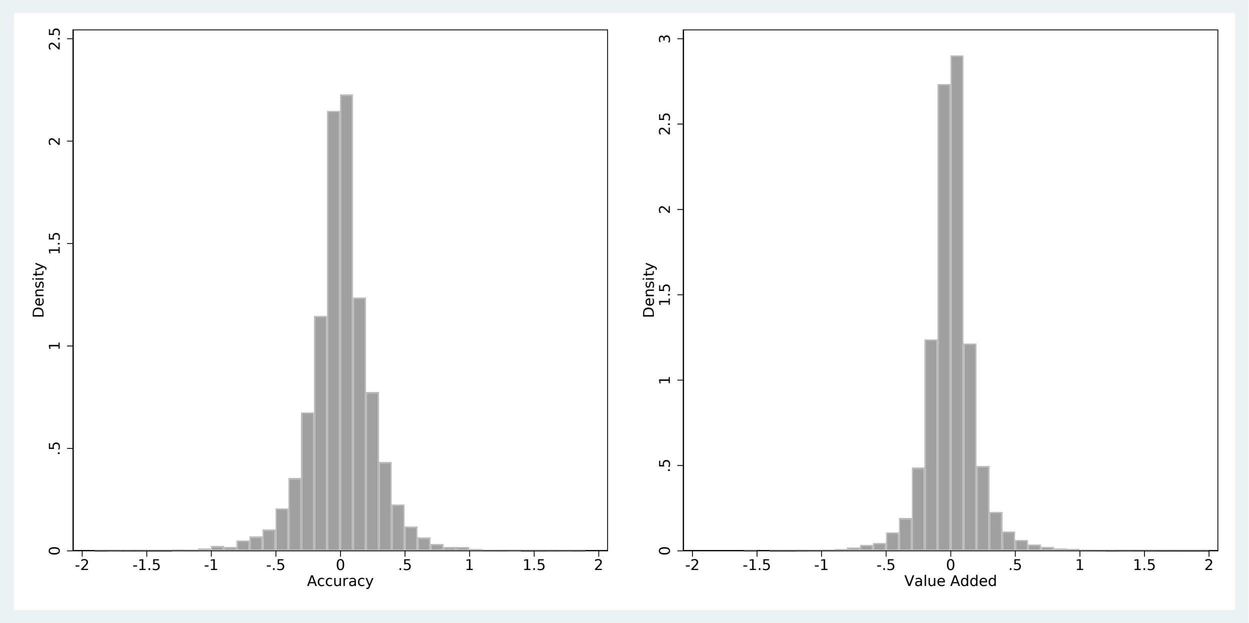
(a) English

(b) Math

(c) Biology

Notes: This plot shows the raw distribution of the assessment deviations D_{ijst}. For each student i, this measure corresponds to the difference between the average test score of all students rated by the teacher in the same achievement level as i in the same school and year, and student i's test score.

Figure 8: Distribution of the Estimated Teacher FE



(a) Inaccuracy ($\hat{\phi}^{I}_{jt}$)

(b) Value-added ($\hat{\phi}^{VA}_{jt}$)

Notes: This plot shows the distribution of the empirical Bayes estimates of $\hat{\phi}^{I}_{jt}$ and $\hat{\phi}^{VA}_{jt}$. See section 4 for specific details about the estimation.

Table 7: Correlation of Empirical Bayes Estimates

	Test Scores ($\hat{\phi}_j^{VA}$)	Inaccuracy ($\hat{\phi}_j^{I}$)	Underassess ($\hat{\phi}_j^{U}$)	Overassess ($\hat{\phi}_j^{O}$)
Test Scores ($\hat{\phi}_j^{VA}$)	1			
Inaccuracy ($\hat{\phi}_j^{I}$)	-0.26	1		
Underassess ($\hat{\phi}_j^{U}$)	0.13	-0.81	1	
Overassess ($\hat{\phi}_j^{O}$)	-0.18	0.76	-0.47	1

Notes: This matrix reports the correlation between the empirical Bayes estimates $\hat{\phi}_{jt}$, using the pooled leave-year-out estimates.

Table 8: Estimated Standard Deviations of Empirical Bayes Estimates

	Test Scores ($\hat{\phi}_j^{VA}$)	Inaccuracy ($\hat{\phi}_j^{I}$)	Underassess ($\hat{\phi}_j^{U}$)	Overassess ($\hat{\phi}_j^{O}$)
Math	0.049	0.086	0.042	0.037
English	0.014	0.094	0.049	0.032
Biology	0.033	0.081	0.042	0.023

Notes: This table reports the estimated standard deviation of each empirical Bayes estimate, separately by subject. Each estimated standard deviation corresponds to the square root of the estimated covariance in mean residuals from equation (2.2) across classrooms for the same teacher. See Appendix A.1 for details.

Table 9: Gender and Racial Differences in Assessments - By Subject

	Dependent Variable: $\overline{\theta}^{T}_{jst} - \theta_{ijst}$								
	Math			English			Biology		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female	0.151***	0.150***	0.150***	0.055***	0.094***	0.095***	0.074***	0.046***	0.041***
	(0.004)	(0.004)	(0.004)	(0.002)	(0.002)	(0.003)	(0.006)	(0.006)	(0.007)
Black	0.001	0.013**	0.011**	-0.090***	-0.038***	-0.040***	-0.094***	-0.043***	-0.048***
	(0.005)	(0.005)	(0.006)	(0.003)	(0.003)	(0.003)	(0.009)	(0.009)	(0.010)
Hispanic	0.011	0.020***	0.018**	-0.082***	-0.049***	-0.051***	-0.081***	-0.038***	-0.042***
	(0.007)	(0.007)	(0.008)	(0.004)	(0.004)	(0.004)	(0.011)	(0.011)	(0.011)
Asian	0.166***	0.132***	0.133***	0.098***	0.045***	0.045***	0.063***	0.040***	0.040***
	(0.014)	(0.014)	(0.014)	(0.006)	(0.006)	(0.006)	(0.010)	(0.010)	(0.010)
Other	-0.012	0.005	0.006	-0.039***	-0.010**	-0.009**	-0.051***	-0.027**	-0.026**
	(0.010)	(0.010)	(0.010)	(0.005)	(0.005)	(0.005)	(0.012)	(0.012)	(0.012)
Teacher FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Teacher-Student Match	No	No	Yes	No	No	Yes	No	No	Yes
Observations	134203	133153	133153	388051	385225	385225	43327	40628	40628
R^2	0.60	0.62	0.62	0.55	0.59	0.59	0.57	0.60	0.60

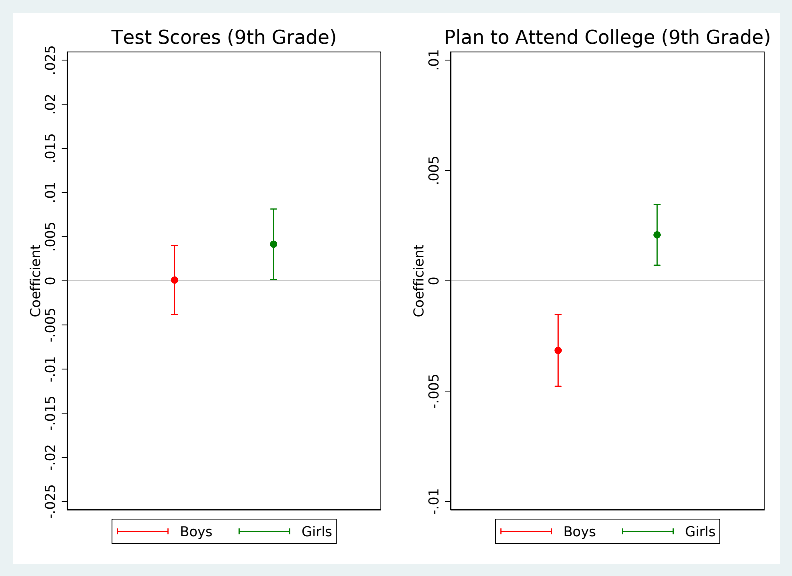
Notes: Each column shows the result of regressing the student-level bias on the student's minority and female status, an indicator equals to one if the student and the teacher belong to the minority group, and the additional controls indicated. Each regression includes classroom and teacher fixed effects, 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: Outcomes in 9th Grade: Main Specification

	Test Score		Plans to Attend College	
	(1)	(2)	(3)	(4)
Value-Added (ϕ_j^{VA})	0.0419***	0.0419***	-0.0003	-0.0003
	(0.0029)	(0.0029)	(0.0006)	(0.0006)
Inaccuracy (ϕ_j^{I})	0.0001	0.0018	-0.0032***	-0.0011*
	(0.0020)	(0.0019)	(0.0008)	(0.0007)
Inaccuracy × Girl	0.0041**		0.0052***	
	(0.0017)		(0.0011)	
Inaccuracy × Black		0.0037		0.0022*
		(0.0023)		(0.0012)
Inaccuracy × Hispanic		-0.0031*		-0.0009
		(0.0030)		(0.0023)
Inaccuracy × Asian		0.0039		0.0005
		(0.0056)		(0.0028)
Inaccuracy × Other		-0.0087**		0.0050**
		(0.0039)		(0.0025)
Mean Dep Var	0.1954	0.1954	0.7819	0.7819
Fixed Effects	Y	Y	Y	Y
Observations	688799	688799	572942	572942
R^2	0.66	0.66	0.14	0.14

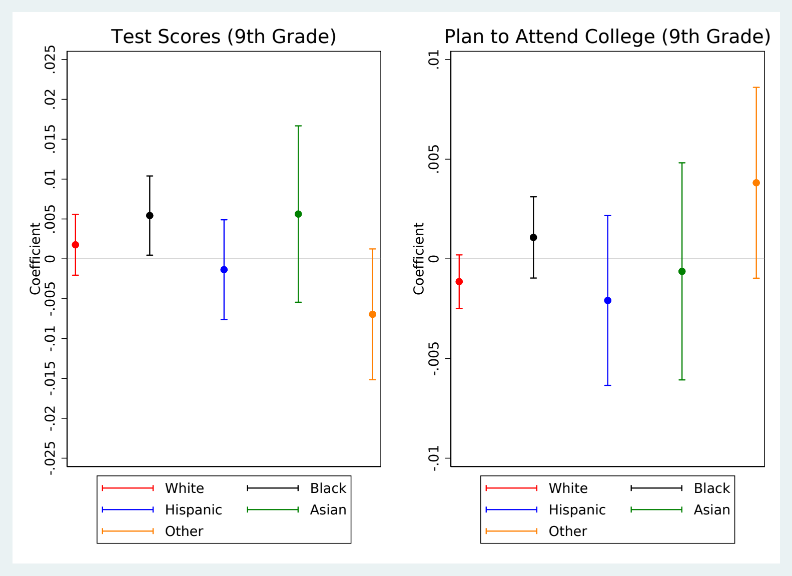
Notes: Clustered standard errors at the teacher and student level in parentheses. Each regression includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and English test scores in 8th grade and 7th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and leave-one-out average classroom characteristics (share of students by race and gender, average math and English scores in 8th grade, share of students by economic disadvantage status, average number of absences, suspensions in 8th grade and 7th grade). All regressions include a set of school-track, year, and subject (math, english, biology) fixed effects. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Figure 9: 9th Grade Outcomes: Coefficients by Gender



Notes: Each subplot shows the estimate of $\hat{\phi}_j^I$ and its 95% confidence interval separately by student gender, based on the results displayed in Table 10. The point estimate and the confidence interval for girls is obtained using the linear combination of the estimate of Inaccuracy (ϕ_j^I) and Inaccuracy × Girl.

Figure 10: 9th Grade Outcomes: Coefficients by Ethnicity



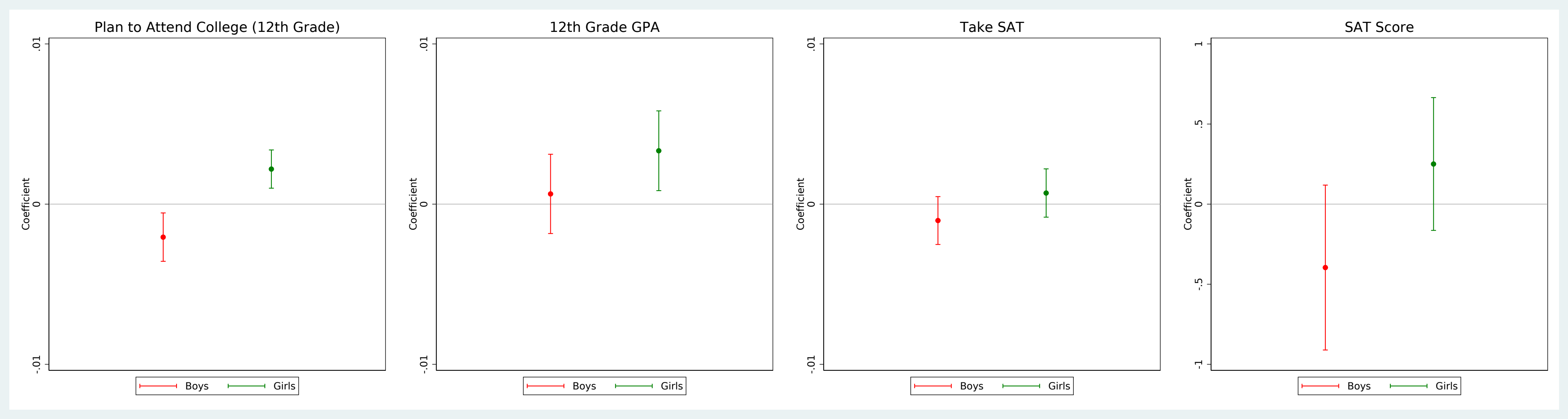
Notes: Each subplot shows the estimate of $\hat{\phi}_j^I$ and its 95% confidence interval separately by student race-ethnicity, based on the results displayed in Table 10. The point estimate and the confidence interval for each subgroup (other than whites) is obtained using the linear combination of the estimate of Inaccuracy (ϕ_j^I) and the interaction of Inaccuracy with the corresponding subgroup.

Table 11: Outcomes in 12th Grade: Main Specification

	GPA 12th		Plans to Attend College		SAT Taker		SAT Scores	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Value-Added (ϕ_j^{VA})	0.0016	0.0016	-0.0000	-0.0000	0.0009	0.0009	-0.1076	-0.0870
	(0.0013)	(0.0013)	(0.0006)	(0.0006)	(0.0008)	(0.0008)	(0.2264)	(0.2246)
Inaccuracy (ϕ_j^{I})	0.0006	-0.0002	-0.0021***	0.0003	-0.0010	-0.0015**	-0.3961	-0.1111
	(0.0013)	(0.0011)	(0.0008)	(0.0006)	(0.0008)	(0.0008)	(0.2626)	(0.2242)
Inaccuracy x Girl	0.0027*		0.0042***		0.0017*		0.6465**	
	(0.0015)		(0.0010)		(0.0010)		(0.2963)	
Inaccuracy x Black		0.0057***		-0.0005		0.0031**		0.7108*
		(0.0021)		(0.0010)		(0.0014)		(0.3786)
Inaccuracy x Hispanic		0.0066*		-0.0014		0.0037*		1.5319**
		(0.0035)		(0.0020)		(0.0021)		(0.6727)
Inaccuracy x Asian		0.0184***		0.0047**		0.0079**		-3.5891***
		(0.0049)		(0.0020)		(0.0034)		(1.1114)
Inaccuracy x Other		-0.0021		-0.0018		0.0028		-1.0413
		(0.0035)		(0.0023)		(0.0023)		(0.8037)
Mean Dep Var	3.1203	3.1203	0.8664	0.8664	0.4603	0.4603	992.1	992.1
Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	531293	531293	574895	574895	688799	688799	317168	317168
R^2	0.72	0.72	0.13	0.13	0.36	0.36	0.81	0.81

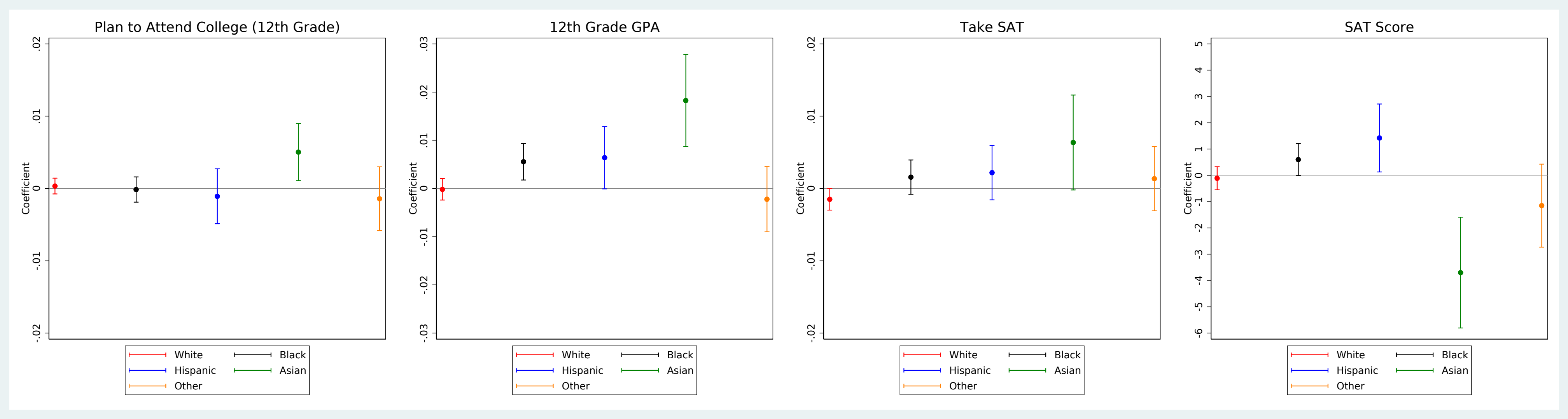
Notes: Clustered standard errors at the teacher level in parentheses. Each regression also includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and English test scores in 8th grade and 7th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and leave-one-out average classroom characteristics (share of students by race and gender, average math and English scores in 8th grade, share of students by economic disadvantage status, average number of absences, suspensions in 8th grade and 7th grade). All regressions include a set of school-track, year, and subject (math, english, biology) fixed effects. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Figure 11: 12th Grade Outcomes: Coefficients by Gender



Notes: Each subplot shows the estimate of $\hat{\phi}_j^I$ and its 95% confidence interval separately by student gender, based on the results displayed in Table 11. The point estimate and the confidence interval for girls is obtained using the linear combination of the estimate of Inaccuracy (ϕ_j^I) and Inaccuracy × Girl.

Figure 12: 12th Grade Outcomes: Coefficients by Ethnicity



Notes: Each subplot shows the estimate of $\hat{\phi}_j^I$ and its 95% confidence interval separately by student race-ethnicity, based on the results displayed in Table 11. The point estimate and the confidence interval for each subgroup (other than whites) is obtained using the linear combination of the estimate of Inaccuracy (ϕ_j^I) and the interaction of Inaccuracy with the corresponding subgroup.

Table 12: Changes in Explained Variance for Different Outcomes

	Value-Added (ϕ_j^{VA}) + Inaccuracy (ϕ_j^{I})	Value-Added (ϕ_j^{VA}) + Underassess (ϕ_j^{U})	Value-Added (ϕ_j^{VA}) + Overassess (ϕ_j^{O})
	(1)	(2)	(3)
Test Scores (9th Grade)	0.2%	1.7%	0.3%
Attends 10th Grade	5%	15%	0.1%
GPA (10th Grade)	-0.4%	4%	33%
Graduation	0.5%	7%	17%
GPA (12th Grade)	209%	278%	77%
Plans to attend college	54%	5%	64%
Take SAT	3%	11%	40%
SAT Score	13%	15%	24%

Notes: This table shows the change in the variance explained by teacher effects, relative to the baseline model (4.4), but using only teacher test score value-added (ϕ_j^{VA}).

Table 13: Selection on Observables Test

	Test Score	Plans College (9th Grade)	In 10th	10th GPA	Graduated	Plans College (12th Grade)	12th GPA	SAT Taker	SAT Score
Value-Added (ϕ_j^{VA})	-0.0005	-0.0001	-0.0000	-0.0005	-0.0000	-0.0002**	-0.0003	-0.0001	0.1409
	(0.0007)	(0.0001)	(0.0001)	(0.0006)	(0.0002)	(0.0001)	(0.0008)	(0.0003)	(0.2096)
Inaccuracy (ϕ_j^I)	0.0000	0.0001	0.0000	0.0001	-0.0000	0.0002	0.0007	-0.0000	-0.0164
	(0.0006)	(0.0001)	(0.0001)	(0.0005)	(0.0001)	(0.0001)	(0.0007)	(0.0002)	(0.1929)
Observations	684911	570942	684911	632861	684911	572220	528858	684911	316388
R^2	0.78	0.44	0.57	0.73	0.49	0.45	0.77	0.66	0.79

Notes: Clustered standard errors at the teacher and student level in parentheses. Each column corresponds to a regression where the left-hand side variable is the predicted outcome using predetermined characteristics and outcomes observed in 7th grade (parental education, gender, race, a third-degree polynomial of math and English test scores in 7th grade, number of suspensions, absences, and repeater status in 7th grade) and the explanatory variables are the teacher characteristics (ϕ_j) and 8th grade controls (a third-degree polynomial of math and English test scores in 8th grade, number of suspensions, absences, and repeater status in 8th grade, 8th grade GPA, and leave-one-out average classroom characteristics). Each regression also includes school-track, year, and subject fixed effects. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 14: Outcomes in 9th Grade: IV Specification

	Test Score		Plans to Attend College	
	(1)	(2)	(3)	(4)
Value-Added (ϕ_j^{VA})	0.0265***	0.0266***	-0.0010	-0.0011
	(0.0043)	(0.0043)	(0.0021)	(0.0021)
Inaccuracy (ϕ_j^I)	-0.0011	0.0003	-0.0080***	-0.0040**
	(0.0037)	(0.0035)	(0.0021)	(0.0019)
Inaccuracy × Girl	0.0048**		0.0111***	
	(0.0023)		(0.0020)	
Inaccuracy × Black		0.0027		0.0045*
		(0.0033)		(0.0025)
Inaccuracy × Hispanic		0.0044		0.0011
		(0.0043)		(0.0045)
Inaccuracy × Asian		0.0159**		0.0013
		(0.0072)		(0.0056)
Inaccuracy × Other		-0.0047		0.0111**
		(0.0052)		(0.0047)
Mean Dep Var	0.1954	0.1954	0.7819	0.7819
F-Test	16194	7946	10720	5319
Fixed Effects	Y	Y	Y	Y
Observations	688527	688527	572753	572753

Notes: Clustered standard errors at the teacher and student level in parentheses. Each column corresponds to an IV regression where the teacher estimate is instrumented with its school-year average. Each regression includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and English test scores in 8th grade and 7th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and leave-one-out average classroom characteristics (share of students by race and gender, average math and English scores in 8th grade, share of students by economic disadvantage status, average number of absences, suspensions in 8th grade and 7th grade). All regressions include a set of school-track, year, and subject (math, english, biology) fixed effects. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 15: Outcomes in 12th Grade: IV Specification

	GPA 12th		Plans to Attend College		SAT Taker		SAT Scores	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Value-Added (ϕ_j^{VA})	0.0054	0.0053	0.0040**	0.0040**	0.0036	0.0035	0.9711	1.0040
	(0.0049)	(0.0049)	(0.0019)	(0.0019)	(0.0028)	(0.0028)	(0.7290)	(0.7232)
Inaccuracy (ϕ_j^I)	-0.0007	-0.0016	-0.0062***	-0.0005	-0.0053**	-0.0071***	0.1833	0.5318
	(0.0042)	(0.0040)	(0.0022)	(0.0018)	(0.0023)	(0.0025)	(0.7091)	(0.7020)
Inaccuracy x Girl	0.0094***		0.0111***		0.0048**		1.1245**	
	(0.0029)		(0.0019)		(0.0019)		(0.5549)	
Inaccuracy x Black		0.0140***		-0.0002		0.0109***		1.9201**
		(0.0049)		(0.0022)		(0.0031)		(0.7906)
Inaccuracy x Hispanic		0.0204***		-0.0033		0.0061		3.0208**
		(0.0072)		(0.0037)		(0.0045)		(1.3633)
Inaccuracy x Asian		0.0322***		0.0144***		0.0225***		-9.7821***
		(0.0104)		(0.0041)		(0.0075)		(2.4605)
Inaccuracy x Other		0.0042		0.0015		0.0126**		-2.8737*
		(0.0073)		(0.0043)		(0.0049)		(1.5429)
Mean Dep Var	3.1203	3.1203	0.8664	0.8664	0.4603	0.4603	992.1	992.1
F-Test	10759	5423	12053	6079	16194	7946	5313	2710
Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	531126	531126	574722	574722	688527	688527	317078	317078

Notes: Clustered standard errors at the teacher and student level in parentheses. Each column corresponds to an IV regression where the teacher estimate is instrumented with its school-year average. Each regression includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and English test scores in 8th grade and 7th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and leave-one-out average classroom characteristics (share of students by race and gender, average math and English scores in 8th grade, share of students by economic disadvantage status, average number of absences, suspensions in 8th grade and 7th grade). All regressions include a set of school-track, year, and subject (math, english, biology) fixed effects. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 16: Outcomes in 9th Grade: Including Teacher Effects in a Different Subject

	Test Score		Plans to Attend College	
	(1)	(2)	(3)	(4)
Value-Added (ϕ_j^{VA})	0.0391***	0.0391***	-0.0008	-0.0008
	(0.0032)	(0.0032)	(0.0011)	(0.0011)
Inaccuracy (ϕ_j^I)	-0.0015	0.0008	-0.0041***	-0.0023**
	(0.0021)	(0.0020)	(0.0012)	(0.0011)
Inaccuracy × Girl	0.0065***		0.0055***	
	(0.0024)		(0.0014)	
Inaccuracy × Black		0.0059**		0.0032*
		(0.0024)		(0.0018)
Inaccuracy × Hispanic		-0.0040		-0.0000
		(0.0036)		(0.0030)
Inaccuracy × Asian		0.0047		-0.0009
		(0.0079)		(0.0034)
Inaccuracy × Other		-0.0052		0.0089**
		(0.0043)		(0.0035)
School-Track F.E.	Y	Y	Y	Y
Year F.E.	Y	Y	Y	Y
Other-Subject Teacher F.E.	Y	Y	Y	Y
Observations	423049	423049	378727	378727
R^2	0.65	0.65	0.15	0.15

Notes: Clustered standard errors at the teacher and student level in parentheses. These regressions are based on a sub-sample of students linked to more than one teacher and include a set of other-subject teacher fixed effects in addition to the set of school-track and year fixed effects included in equation (4.4). Each regression includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and English test scores in 8th grade and 7th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and leave-one-out average classroom characteristics (share of students by race and gender, average math and English scores in 8th grade, share of students by economic disadvantage status, average number of absences, suspensions in 8th grade and 7th grade). *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table 17: Outcomes in 12th Grade: Including Teacher Effects in a Different Subject

	GPA 12th		Plans to Attend College		SAT Taker		SAT Scores	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Value-Added (ϕ_j^{VA})	0.0012	0.0011	-0.0008	-0.0008	-0.0017	-0.0017	-0.0894	-0.0815
	(0.0019)	(0.0019)	(0.0010)	(0.0010)	(0.0013)	(0.0013)	(0.3489)	(0.3489)
Inaccuracy (ϕ_j^I)	-0.0024	-0.0018	-0.0018	0.0000	-0.0034***	-0.0041***	-0.0189	-0.3998
	(0.0018)	(0.0018)	(0.0011)	(0.0009)	(0.0012)	(0.0012)	(0.3769)	(0.3548)
Inaccuracy x Girl	0.0060***		0.0045***		0.0039***		0.6013	
	(0.0018)		(0.0013)		(0.0014)		(0.3796)	
Inaccuracy x Black		0.0064**		0.0008		0.0067***		0.0979
		(0.0028)		(0.0014)		(0.0020)		(0.4937)
Inaccuracy x Hispanic		0.0083*		0.0008		0.0086***		1.2885
		(0.0046)		(0.0027)		(0.0028)		(0.9300)
Inaccuracy x Asian		0.0177***		0.0025		0.0046		-3.9584***
		(0.0063)		(0.0027)		(0.0043)		(1.3488)
Inaccuracy x Other		-0.0029		0.0041		0.0076**		0.0151
		(0.0049)		(0.0030)		(0.0034)		(1.0552)
School-Track F.E.	Y	Y	Y	Y	Y	Y	Y	Y
Year F.E.	Y	Y	Y	Y	Y	Y	Y	Y
Other-Subject Teacher F.E.	Y	Y	Y	Y	Y	Y	Y	Y
Observations	337369	337369	360268	360268	423049	423049	206772	206772
R^2	0.71	0.71	0.16	0.16	0.36	0.36	0.82	0.82

Notes: Clustered standard errors at the teacher and student level in parentheses. These regressions are based on a sub-sample of students linked to more than one teacher and include a set of other-subject teacher fixed effects in addition to the set of school-track and year fixed effects included in equation (4.4). Each regression includes individual controls at the student level (parental education, gender, race, a third-degree polynomial of math and English test scores in 8th grade and 7th grade, number of suspensions, absences, and repeater status in 8th grade and 7th grade), and leave-one-out average classroom characteristics (share of students by race and gender, average math and English scores in 8th grade, share of students by economic disadvantage status, average number of absences, suspensions in 8th grade and 7th grade). *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

A Appendix

A.1 Computation of the Variance Terms

Based on 4.1, I assume the following structure of the residual term $\varepsilon_{ijst} = D_{ijst} - X'_{ist}\hat{\gamma} - C'_{ijst}\hat{\delta} - \hat{\tau}_s$:

$$\varepsilon_{ijst} = \phi_j + \varepsilon_{ijstc} + \xi_{ijst}$$

To estimate the variance of each term, I follow Kane and Staiger (2008) and impose that the cross-covariances are zero. Using this assumption we can write:

$$\sigma^2_{\varepsilon_{ijst}} = \sigma^2_{\phi} + \sigma^2_{\varepsilon_{ijstc}} + \sigma^2_{\xi_{ijst}}$$

Where σ^2_{ϕ} corresponds to the variance of teacher effects, $\sigma^2_{\varepsilon_{ijstc}}$ to the variance of classroom shocks, and $\sigma^2_{\xi_{ijst}}$ to the variance of student-specific shocks. The variance of teacher effects can be estimated using the covariance of the average residuals at the classroom level, across different years:

$$\text{cov}(\bar{\varepsilon}_{jst}, \bar{\varepsilon}_{js't'}) = \sigma^2_{\phi}$$

The variance of student-level shocks is estimated by computing the residual variance of the errors from a regression of the residuals onto a full set of classroom fixed effects:

$$\varepsilon_{ijst} = \alpha + \varphi_c + \tilde{\xi}_{ijst}$$

Finally, the covariance of classroom-level shocks corresponds to the difference between the variance of the residual and the two variances described above.

$$\hat{\sigma}^2_{\varepsilon_{ijstc}} = \hat{\sigma}^2_{\varepsilon_{ijst}} - \hat{\sigma}^2_{\phi} - \hat{\sigma}^2_{\xi_{ijst}}$$

Following Jackson (2018), the estimate of the variance of teacher effects ($\hat{\sigma}^2_{\phi}$) is calculated using a bootstrap procedure. I compute 500 covariance estimates $\text{cov}(\bar{\varepsilon}_{jst}, \bar{\varepsilon}_{js't'})$ by randomly pairing classrooms in different years for the same teacher. I employ the median value as the parameter estimate.

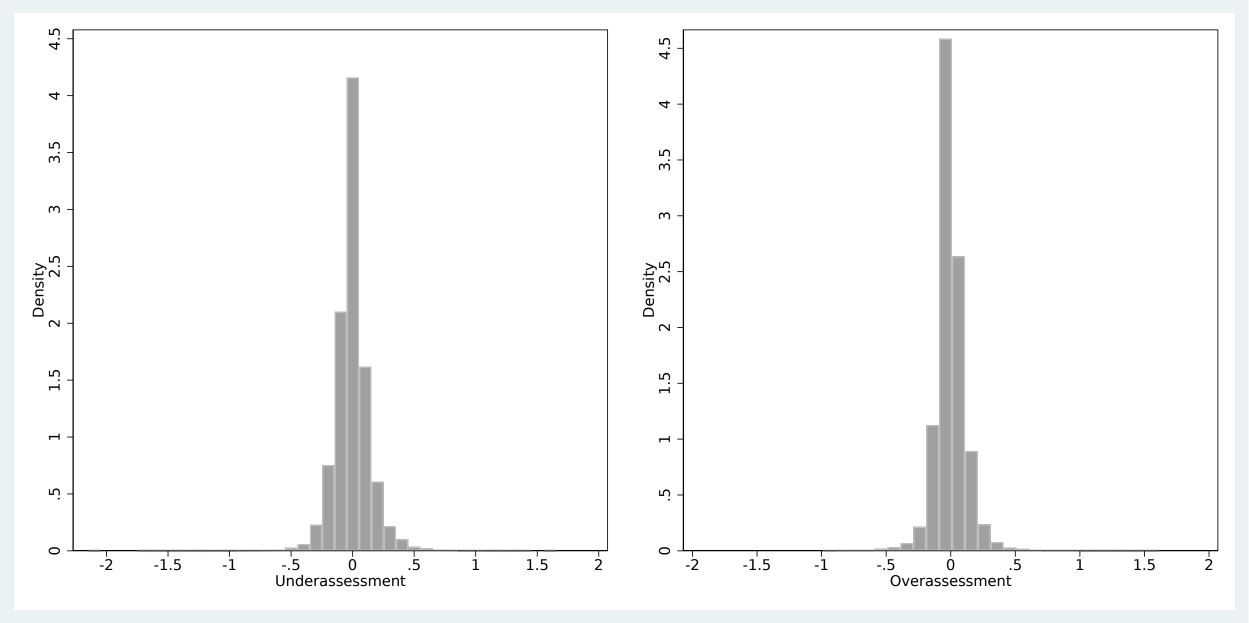
A.2 Additional Figures and Tables

Table A1: Gender and Racial Differences in Assessments - By Subject

	Dependent Variable: $\overline{\theta}^{T}_{jst} - \theta_{ijst}$		
	Math	English	Biology
	(1)	(2)	(3)
Female	0.140***	0.090***	0.035***
	(0.004)	(0.003)	(0.008)
Black	0.010*	-0.039***	-0.045***
	(0.006)	(0.003)	(0.012)
Hispanic	0.018**	-0.048***	-0.051***
	(0.008)	(0.004)	(0.013)
Asian	0.126***	0.043***	0.035***
	(0.014)	(0.006)	(0.012)
Other	0.005	-0.010**	-0.033**
	(0.010)	(0.005)	(0.014)
Assessment 8th grade	0.121***	0.080***	0.070***
	(0.004)	(0.002)	(0.006)
Teacher FE	Yes	Yes	Yes
Classroom FE	Yes	Yes	Yes
Student Controls	Yes	Yes	Yes
Teacher-Student Match	Yes	Yes	Yes
Observations	130396	381973	27973
R^2	0.63	0.59	0.60

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.

Figure A1: Distribution of the Estimated Teacher FE



(a) Underassessment ($\hat{\phi}^{U}_{jt}$)

(b) Overassessment ($\hat{\phi}^{O}_{jt}$)

Notes: This plot shows the distribution of the empirical Bayes estimates of $\hat{\phi}^{I}_{jt}$ and $\hat{\phi}^{VA}_{jt}$. See section 4 for specific details about the estimation.

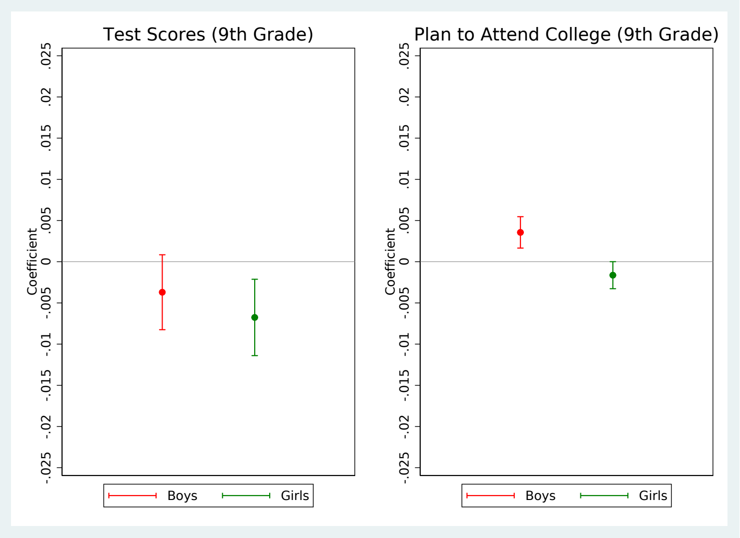
Table A2: Selection on Observables Test: Additional Teacher Measures

	Test Score	Plans College (9th Grade)	In 10th	10th GPA	Graduated	Plans College (12th Grade)	12th GPA	SAT Taker	SAT Score
Value-Added (ϕ_j^{VA})	-0.0006	-0.0001	0.0000	-0.0007	-0.0000	-0.0002**	-0.0007	-0.0003	-0.0245
	(0.0005)	(0.0001)	(0.0001)	(0.0005)	(0.0001)	(0.0001)	(0.0007)	(0.0002)	(0.1771)
Underassess (ϕ_j^{U})	0.0009*	-0.0000	0.0000	0.0006	-0.0001	-0.0000	0.0005	0.0002	0.1942
	(0.0006)	(0.0001)	(0.0001)	(0.0005)	(0.0001)	(0.0001)	(0.0007)	(0.0002)	(0.1696)
Observations	688776	572942	688776	635615	688776	574877	531275	688776	317163
R^2	0.78	0.44	0.57	0.73	0.49	0.44	0.76	0.66	0.79

	Test Score	Plans College (9th Grade)	In 10th	10th GPA	Graduated	Plans College (12th Grade)	12th GPA	SAT Taker	SAT Score
Value-Added (ϕ_j^{VA})	-0.0005	-0.0000	0.0000	-0.0005	0.0000	-0.0002*	-0.0004	-0.0002	0.0310
	(0.0006)	(0.0001)	(0.0001)	(0.0005)	(0.0001)	(0.0001)	(0.0007)	(0.0002)	(0.1778)
Overassess (ϕ_j^{O})	0.0003	0.0002	0.0001	0.0007	0.0001	0.0004	0.0018*	0.0001	0.1753
	(0.0008)	(0.0002)	(0.0002)	(0.0007)	(0.0002)	(0.0003)	(0.0009)	(0.0003)	(0.2840)
Observations	688799	572942	688799	635636	688799	574895	531293	688799	317168
R^2	0.78	0.44	0.57	0.73	0.49	0.44	0.76	0.66	0.79

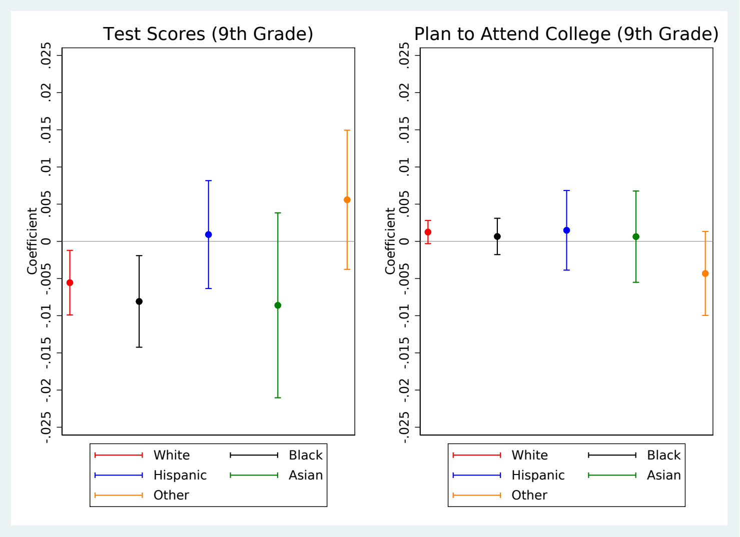
Notes: Clustered standard errors at the teacher and student level in parentheses. Each column corresponds to a regression where the left-hand side variable is the predicted outcome using predetermined characteristics and outcomes observed in 7th grade (parental education, gender, race, a third-degree polynomial of math and English test scores in 7th grade, number of suspensions, absences, and repeater status in 7th grade) and the explanatory variables are the teacher characteristics (ϕ_j) and 8th grade controls (a third-degree polynomial of math and English test scores in 8th grade, number of suspensions, absences, and repeater status in 8th grade, 8th grade GPA, and leave-one-out average classroom characteristics). Each regression also includes school-track, year, and subject (math, english, biology) fixed effects. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Figure A2: Underassessment and 9th Grade Outcomes: Coefficients by Gender



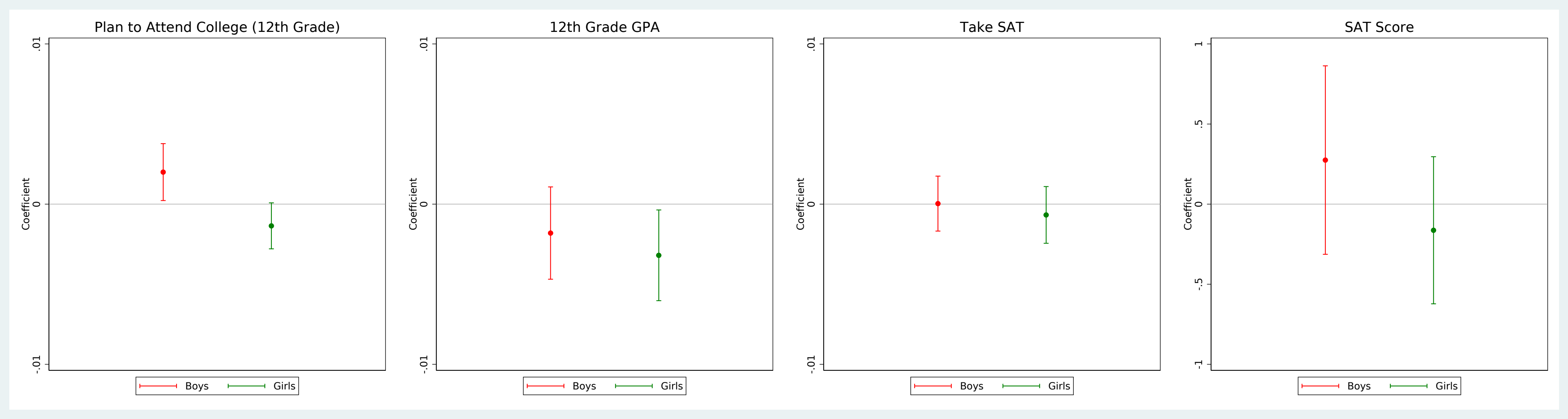
Notes: Each subplot shows the estimate of $\hat{\phi}_j^U$ and its 95% confidence interval separately by student gender, based on the estimates of equation (4.4) employing $\hat{\phi}_j^U$ as the regressor of interest. The point estimate and the confidence interval for girls is obtained using the linear combination of the estimate of underassessment (ϕ_j^U) and underassessment × Girl.

Figure A3: Underassessment and 9th Grade Outcomes: Coefficients by Ethnicity



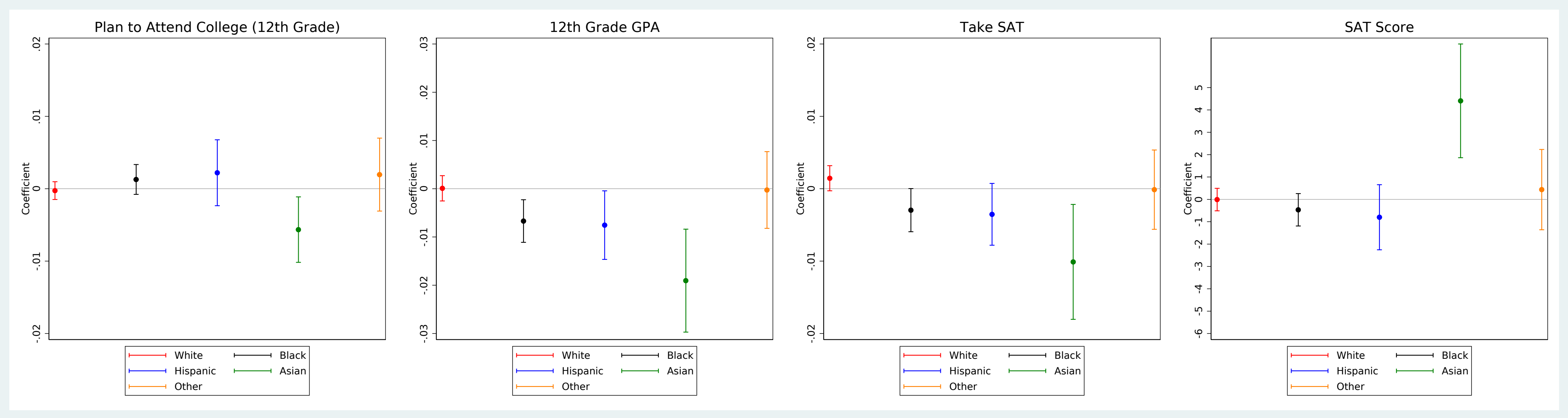
Notes: Each subplot shows the estimate of $\hat{\phi}_j^U$ and its 95% confidence interval separately by student race-ethnicity, based on the estimates of equation (4.4) employing $\hat{\phi}_j^U$ as the regressor of interest. The point estimate and the confidence interval for each subgroup (other than whites) is obtained using the linear combination of the estimate of underassessment (ϕ_j^U) and the interaction of underassessment with the corresponding subgroup.

Figure A4: Underassessment and 12th Grade Outcomes: Coefficients by Gender



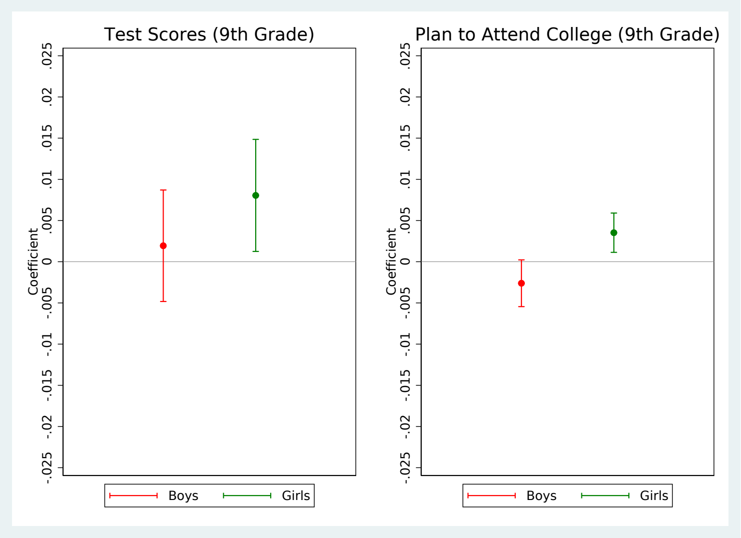
Notes: Each subplot shows the estimate of $\hat{\phi}_j^U$ and its 95% confidence interval separately by student gender, based on the estimates of equation (4.4) employing $\hat{\phi}_j^U$ as the regressor of interest. The point estimate and the confidence interval for girls is obtained using the linear combination of the estimate of underassessment (ϕ_j^U) and underassessment × Girl.

Figure A5: Underassessment and 12th Grade Outcomes: Coefficients by Ethnicity



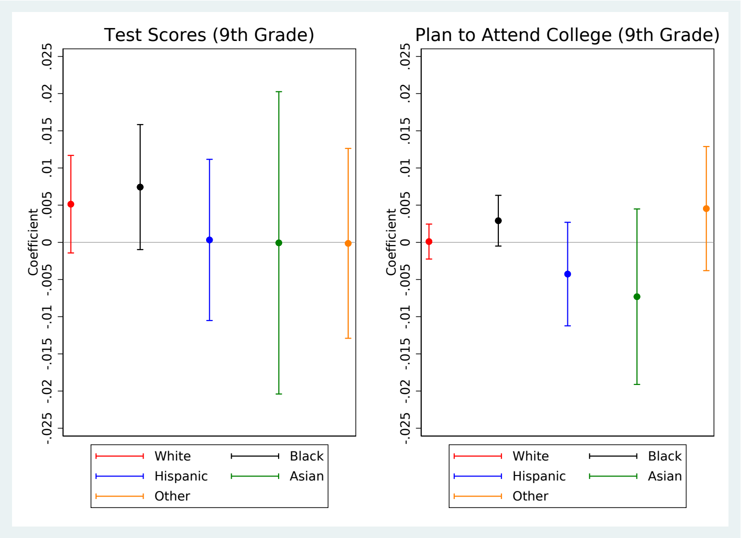
Notes: Each subplot shows the estimate of $\hat{\phi}_j^U$ and its 95% confidence interval separately by student race-ethnicity, based on the estimates of equation (4.4) employing $\hat{\phi}_j^U$ as the regressor of interest. The point estimate and the confidence interval for each subgroup (other than whites) is obtained using the linear combination of the estimate of underassessment (ϕ_j^U) and the interaction of underassessment with the corresponding subgroup.

Figure A6: Overassessment and 9th Grade Outcomes: Coefficients by Gender



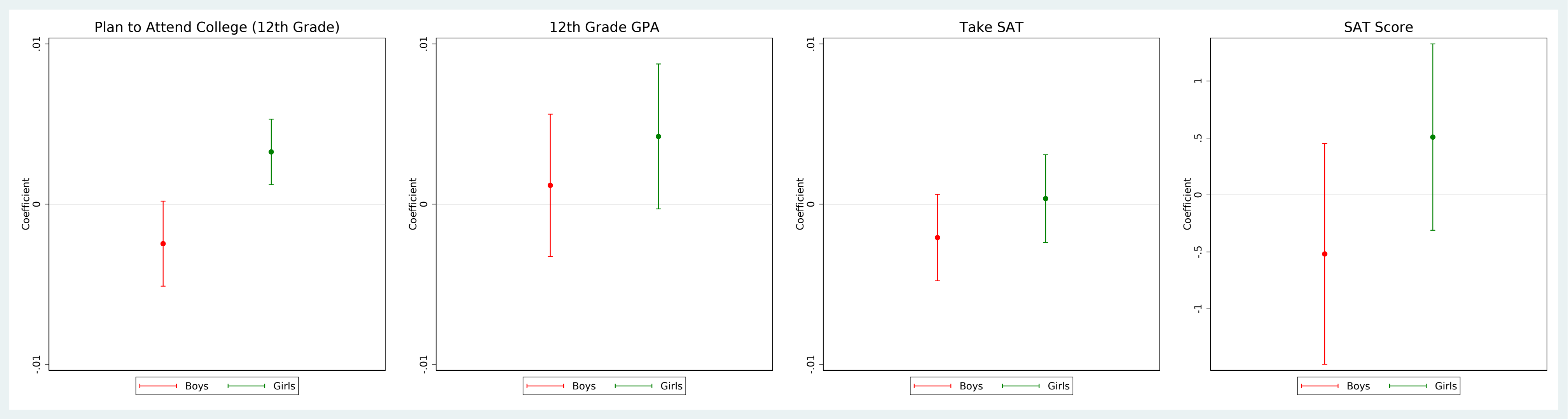
Notes: Each subplot shows the estimate of $\hat{\phi}_j^O$ and its 95% confidence interval separately by student gender, based on the estimates of equation (4.4) employing $\hat{\phi}_j^O$ as the regressor of interest. The point estimate and the confidence interval for girls is obtained using the linear combination of the estimate of overassessment (ϕ_j^O) and overassessment × Girl.

Figure A7: Overassessment and 9th Grade Outcomes: Coefficients by Ethnicity



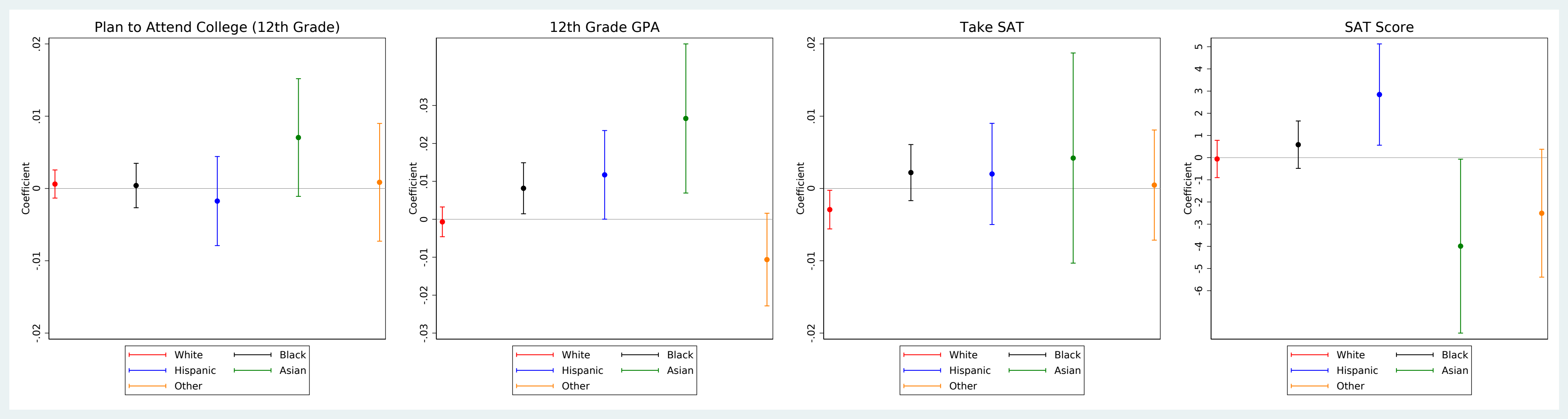
Notes: Each subplot shows the estimate of $\hat{\phi}_j^O$ and its 95% confidence interval separately by student race-ethnicity, based on the estimates of equation (4.4) employing $\hat{\phi}_j^O$ as the regressor of interest. The point estimate and the confidence interval for each subgroup (other than whites) is obtained using the linear combination of the estimate of overassessment (ϕ_j^O) and the interaction of overassessment with the corresponding subgroup.

Figure A8: Overassessment and 12th Grade Outcomes: Coefficients by Gender



Notes: Each subplot shows the estimate of $\hat{\phi}_j^O$ and its 95% confidence interval separately by student gender, based on the estimates of equation (4.4) employing $\hat{\phi}_j^O$ as the regressor of interest. The point estimate and the confidence interval for girls is obtained using the linear combination of the estimate of overassessment (ϕ_j^O) and overassessment × Girl.

Figure A9: Overassessment and 12th Grade Outcomes: Coefficients by Ethnicity



Notes: Each subplot shows the estimate of $\hat{\phi}_j^O$ and its 95% confidence interval separately by student race-ethnicity, based on the estimates of equation (4.4) employing $\hat{\phi}_j^O$ as the regressor of interest. The point estimate and the confidence interval for each subgroup (other than whites) is obtained using the linear combination of the estimate of overassessment (ϕ_j^O) and the interaction of overassessment with the corresponding subgroup.