Zhu, Maria

Working Paper

New Findings on Racial Bias in Teachers' Evaluations of Student Achievement

IZA Discussion Papers, No. 16815

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Zhu, Maria (2024) : New Findings on Racial Bias in Teachers' Evaluations of Student Achievement, IZA Discussion Papers, No. 16815, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/295838

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16815

New Findings on Racial Bias in Teachers' Evaluations of Student Achievement

Maria Zhu

FEBRUARY 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16815

New Findings on Racial Bias in Teachers' Evaluations of Student Achievement

Maria Zhu
Syracuse University and IZA

FEBRUARY 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

New Findings on Racial Bias in Teachers' Evaluations of Student Achievement

This paper examines racial discrepancies in teachers' evaluations of student achievement, conditional on standardized test achievement. After correcting for measurement error in standardized test scores, results indicate teachers evaluate Black students as higher achieving than White students with the same standardized test achievement. This finding stands in contrast to prior findings on Black-White teacher assessment gaps. Further analysis indicates these findings are consistent with two potential explanations: first, standardized tests may exhibit bias against Black students, and second, teachers may inflate assessments of Black students relative to White students due to social desirability bias.

JEL Classification: I20, J15
Keywords: teacher assessments, racial bias, standardized tests

Corresponding author:
Maria Zhu
Syracuse University
900 S Crouse Ave
13244 Syracuse NY
USA
E-mail: mzhu33@syr.edu

I Introduction

A well-established body of evidence finds that teachers' expectations of students have significant effects on academic outcomes.[1] A key concern arising from these findings is that teachers' perceptions of student achievement may be influenced by factors other than student performance, such as negative stereotypes or biases towards certain groups along gender or race lines. This underscores the importance of understanding the nature of biases in student assessment. Existing studies looking at racial discrepancies in teachers' evaluations of student achievement have found that teachers assess Black students as lower performing compared to White students, conditional on standardized test scores (Burgess and Greaves, 2013; Botelho et al., 2015). These findings have been interpreted as evidence of negative bias in teachers' perceptions of Black students.

This paper provides novel evidence on racial biases in student assessments in the US using data on the universe of public school students in North Carolina from 2007-2013. To assess the presence of racial biases in teacher evaluations, I use two measures of achievement. First, students in grades 3-8 in North Carolina take standardized end-of-grade tests in both math and reading. Second, teachers are asked at the end of the year to evaluate their students on a scale of 1 to 4 on their mastery of math and reading along the same skills that standardized tests are designed to evaluate. Following prior studies, I assess whether teachers systematically differ in their assessments of students by race, controlling for standardized test scores. In addition to test score controls, I also include classroom fixed effects to control for unobserved factors varying across classes that may affect evaluations (e.g., teacher-specific assessment standards, course of study). The key identifying assumption is that conditional on underlying standardized test scores, systematic racial differences in teacher assessments within a classroom reflect the effects of race, rather than unobserved characteristics correlated with race.

One concern with this approach is that test score assessments capture a single snapshot of student achievement on one day from a limited number of multiple choice questions. As such, standardized tests measure true ability with noise for a variety of reasons, including

[1] See Rosenthal and Jacobson (1968); Papageorge et al. (2020); Hill and Jones (2021); Lavy and Sand (2018); Lavy and Megalokonomou (2019); Lindahl (2016).

test instrument noise (e.g., randomness in the selection of questions on the test) and variability in testing conditions (e.g., student illness on testing day, how well student slept the night before). I first discuss conceptually how measurement error in standardized tests will bias estimates on racial gaps in teachers' assessments of students in this setting, given the correlation between student achievement and race. Next, I use Monte Carlo simulations to demonstrate that the biases generated by measurement error will lead to conclusions that teachers assess Black students as lower performing than White students with identical standardized test scores, even in the absence of teacher bias. Subsequently, I use Monte Carlo simulations to demonstrate that an instrument that satisfies the assumptions of a) being correlated with test scores and b) affecting teacher assessments only through its correlation with test scores will yield unbiased estimates. Additionally, I discuss five potential candidate instruments and the relative advantages and limitations of each: lagged test scores and twice lagged test scores in the tested subject, other subject test scores, lagged and twice lagged other subject test scores.

Results estimated using ordinary least squares (OLS) that do not correct for measurement error find results that are similar to those from prior studies: teachers evaluate Black students as lower achieving than they do observationally similar White peers with the same standardized test scores.[2] However, results change significantly in this setting after correcting for measurement error. Instrumental variables (IV) results indicate teachers are 1.8-2.7 percentage points more likely to evaluate Black students as grade-level proficient in math achievement than they are White peers with the same standardized test-based achievement, a 2.4-3.6 percent increase from the mean. Similarly, teachers are 0.6 to 3.2 percentage points more likely to evaluate Black students as proficient in reading than they are White peers with the same standardized test-based achievement in math, a 0.8-4.3 percent increase from the mean. Results are robust to tests assessing concerns that findings are driven by unobserved behavioral correlates of race, invalid instruments, entry-level achievement influencing teachers' evaluations, and limited common support.

These findings align with multiple potential underlying mechanisms. One possible explanation is that standardized test score bias against Black students could result in a positive

[2]See Burgess and Greaves (2013) and Botelho et al. (2015).

disparity in teacher assessments for Black students compared to White students achieving the same standardized test results.[3] If teachers do not exhibit bias against black students, or if the bias of teachers is lesser than that present in the tests, teacher evaluations will rate Black students higher than White students with equivalent test performance in the presence of test score bias.

Alternatively, these outcomes might stem from racial disparities in teacher evaluations, rather than in test scores. First, teachers may harbor lower expectations for Black students, leading them to inflate ratings for these students in comparison to equally performing White students in the classroom. In such a scenario, any bias in test scores would exacerbate the positive gap for Black students relative to their White counterparts with identical test scores. Second, teachers might adjust their evaluation criteria to accommodate student backgrounds. If teachers perceive that Black students, on average, face more obstacles to achieve a given level of academic proficiency compared to White students, they may calibrate their assessments accordingly. For instance, a teacher may believe that a student from a disadvantaged background who attains the same academic mastery as a student from a more privileged background demonstrates a higher level of academic aptitude and adjust their assessments to reflect this. Third, these results could be influenced by social desirability bias, wherein teachers inflate their assessments of Black students to align with responses they perceive as more socially favorable.

I conduct multiple analyses to investigate the plausibility of each of these mechanisms. I find that racial gaps in student evaluations do not differ by teacher race. Additionally, teachers consistently rate economically disadvantaged students as lower achieving than their non-economically disadvantaged counterparts, controlling for race. Finally, in counties where the income gap between White and Black populations is wider, teachers tend to assess Black students as relatively lower achieving. Overall, results suggest a potential influence of standardized test bias or social desirability bias. However, without more information on test instruments, I am unable to definitively pin down mechanisms using administrative data.

This paper contributes to our comprehension of group-level disparities in teachers' eval-

[3]Prior research has documented test score bias against students from underrepresented groups (Sandoval, 1979; Freedle, 2003; Duncan and Sandy, 2013).

uations of students. A number of papers have examined racial or gender gaps in teachers' assessments of student achievement (Lavy, 2008; Lindahl, 2016; Botelho et al., 2015; Burgess and Greaves, 2013; Rangel and Shi, 2021; Shi and Zhu, 2023; Francis et al., 2019; Francis, 2012). Of particular relevance to this study, a couple of other papers have investigated differences in teacher assessments of achievement between Black and White students, conditional on standardized test scores. Burgess and Greaves (2013) examines racial differences in teacher assessments in England and finds that teachers are significantly more negative in their assessments of Black Caribbean and Black African students compared to their White counterparts with the same test scores. Using Brazilian data, Botelho et al. (2015) also finds evidence of teacher bias in the form of more negative assessments toward Black students relative to White peers. In contrast, this paper examines racial discrepancies in teachers' assessments of Black and White students in the US, and notably finds very different results. These findings are particularly intriguing as Black students face greater disadvantages compared to White students across all three settings. My results underscore institutional differences across countries that may give rise to varied expressions of bias. Specifically, my findings suggest that racial bias in standardized tests or social pressures faced by teachers overshadow negative teacher bias in evaluations in this context.[4]

This paper also relates to a body of empirical research on bias induced by measurement error. Bound et al. (2001) highlights the wide degree of measurement error present in survey reports of labor-related phenomena. Gillen et al. (2019) replicates three classic economic experiments using survey data to demonstrate that correcting for measurement error can lead to substantially different results. Ward (2023) and Modalsli and Vosters (2022) address the importance of correcting for measurement error in quantifying intergenerational mobility. Modalsli and Vosters (2022) finds that correcting for measurement error substantially affects estimates of multigenerational mobility. Ward (2023) finds that accounting for measurement error and race leads to substantially larger estimates of intergenerational persistence than previously believed. Most closely related to this paper are a smaller subset of studies docu-

[4]Another potential explanation for some of the discrepancies between studies is the role of measurement error.Burgess and Greaves (2013) does not address test score measurement error in their study. Botelho et al. (2015) does correct for measurement error using an IV approach with lagged test scores in the same subject. The paper finds that after correcting for measurement error, teachers still assess Black students as lower achieving than White students, although magnitudes of the gap are reduced by half.

menting the importance of correcting for measurement error in assessing disparities between groups. Hanushek and Rivkin (2009) and Bond and Lang (2017) emphasize the importance of correcting for measurement error in standardized test scores when measuring the evolution of racial achievement gaps over time using administrative data. This study builds upon this body of work by demonstrating the importance of correcting for measurement error in standardized test scores to measure racial disparities in teacher assessments across assessment metrics. Relatedly, van Huizen et al. (2024) documents the importance of correcting for measurement error in teacher track recommendations when examining differences in recommendations by student socioeconomic status. Crucially, I find that estimation results after correcting for measurement error depart significantly from both results that do not correct for measurement error and from prior studies of this question, which have looked at different countries. These novel findings shed light on the nature of racial biases in student assessments, with important implications for policy initiatives targeting racial disparities within this context.

In the remainder of the paper, Section II introduces the data and empirical setting of this study. Section III presents the empirical strategy used to quantify racial gaps in teachers' evaluations of students. Section IV illustrates the effect of measurement error in test scores on estimation and discusses methods of correcting for this concern. Section V presents results, and Section VI concludes the paper.

II Data and Descriptive Statistics

II.A Data Overview

Data for this project come from the North Carolina Education Research Data Center (NCERDC), comprising administrative records for all public school students and teachers in North Carolina. For students, I observe information gender, race and ethnicity, whether the student is economically disadvantaged, number days absent in a year, and detailed disciplinary infraction information. For teachers, I observe information on race and ethnicity, age, and teaching experience. Course membership rosters allow me to link students to their teachers

across classes. My analysis focuses on students in grades 3-8 from 2007-2013.[5]

Key to this study, I also observe two metrics of student achievement. The first measure comes from end-of-grade standardized test scores in math and reading. Standardized tests are multiple choice and machine-scored. In addition to test scores, I observe teacher evaluations of students. These evaluations are collected around the time standardized tests are administered and are obtained before teachers learn standardized test results. Specifically, teachers are asked to evaluate students in math and reading proficiency on a discrete scale of 1 to 4, corresponding to the following qualitative achievement descriptions:

1. **Insufficient mastery**: Students performing at this level do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level.

2. **Inconsistent mastery**: Students performing at this level demonstrate inconsistent mastery of knowledge and skills in this subject area and are minimally prepared to be successful at the next grade level.

3. **Consistent mastery**: Students performing at this level consistently demonstrate mastery of grade level subject matter and skills and are well prepared for the next grade level.

4. **Superior performance**: Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade level work.

An evaluation at level 3 or above indicates the teacher determines the student to have achieved grade-level proficiency or higher in the subject being assessed. Importantly, teachers' evaluations of students carry no direct consequences for either the teacher or the student, thus mitigating incentives for inaccurate reporting. The primary purpose behind the state's collection of teacher assessments is to incorporate them as one of several inputs to refine interpretations of standardized test outcomes. Moreover, teachers receive explicit guidelines on student assessment, ensuring consistency with the criteria measured by standardized tests. These guidelines emphasize the importance of evaluating students based on actual mastery

[5]I use this time frame because this is the set years for which the NCERDC has data on both standardized tests scores and teacher assessments.

of the material, rather than on behavioral traits that may be associated with academic performance. Specifically, teachers for a given subject are instructed:

> "The [subject] teacher should base this response for each student solely on mastery of [subject]. The [subject] teacher may elect to use grades as a starting point in making these assignments. However, grades are often influenced by factors other than pure achievement, such as failure to turn in homework. The [subject] teacher's challenge is to provide information that reflects only the achievement of each student in the subject matter tested."[6]

One important distinction to make is that teacher assessments in this setting focus on judgments regarding current mastery of skills. This differs from assessments of expectations for future success, as examined in Papageorge et al. (2020). Although both assessments of current achievement and expectations for future achievement may exert tangible effects on students—such as influencing a teacher's inclination to recommend a student for honors classes—they do not necessarily gauge the same aspect. For example, it is possible for a teacher to assess two students as possessing equivalent levels of current achievement but anticipate divergent trajectories in terms of future academic pursuits. This study concentrates on comparing two contemporaneous measures of achievement—namely, test scores and teacher assessments of current mastery—while abstracting from differences in factors like student ambition and family background that might influence future achievement.

II.B Descriptive Statistics

The top panel of Table 1 provides information on students in the sample. Approximately 54 percent of students are White, 27 percent are Black, 12 percent are Hispanic, and 8 percent identify as a different racial/ethnic group. About half of students are female and about half of students are classified as economically disadvantaged. The bottom panel of Table 1 provides information on teachers in the sample. A large majority of teachers in a given year are White, at 84 percent, while 14 percent are Black and the remaining 2 percent of teachers identify as a different race/ethnicity. Approximately 89 percent of teachers are female, and

[6] Hill and Jones (2021)

teachers have an average of 11.4 years of experience, as calculated by the number of years they have taught in the North Carolina public school system.

Table 1: Student and Teacher Characteristics

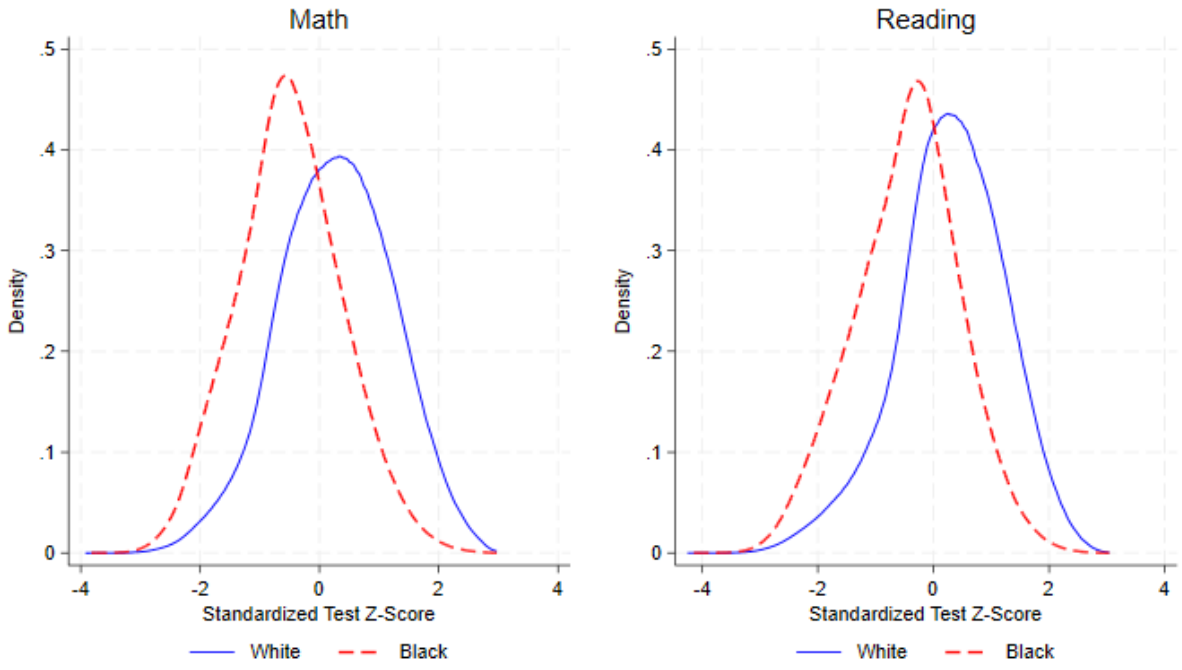
	Mean
Students	
White	0.54
Black	0.27
Hispanic	0.12
Other race	0.08
Female	0.49
Economically disadvantaged	0.51
N	4,609,299
Teachers	
White teacher	0.84
Black teacher	0.14
Other race teacher	0.02
Female teacher	0.89
Teacher experience (years)	11.36
	(9.24)
N	187,548

Observations in the top panel represent the student-year level observations for students in grades 3-8 in math or reading classes between 2007-2013. Observations in the bottom panel represent the teachers-year level observations for individuals teaching grades 3-8 in math or reading classes during this time period.

Next, Figure 1 displays standardized test score distributions in math and reading by student race. Test scores are normalized so that a one-unit increase in test scores represents a one standard deviation change. For both subjects, the distribution of test scores for White students is first order stochastically dominant to the distribution of test scores for Black students for virtually the entire support.

To get a sense for how teacher assessments of students differ by race, Figure 2 depicts teacher assessments by standardized test achievement level. The state discretizes all raw test scores into four levels of achievement. These levels of achievement are meant to qualitatively correspond to the four categories of achievement on which teachers assess students. The

Figure 1: Standardized Test Scores Distributions by Race

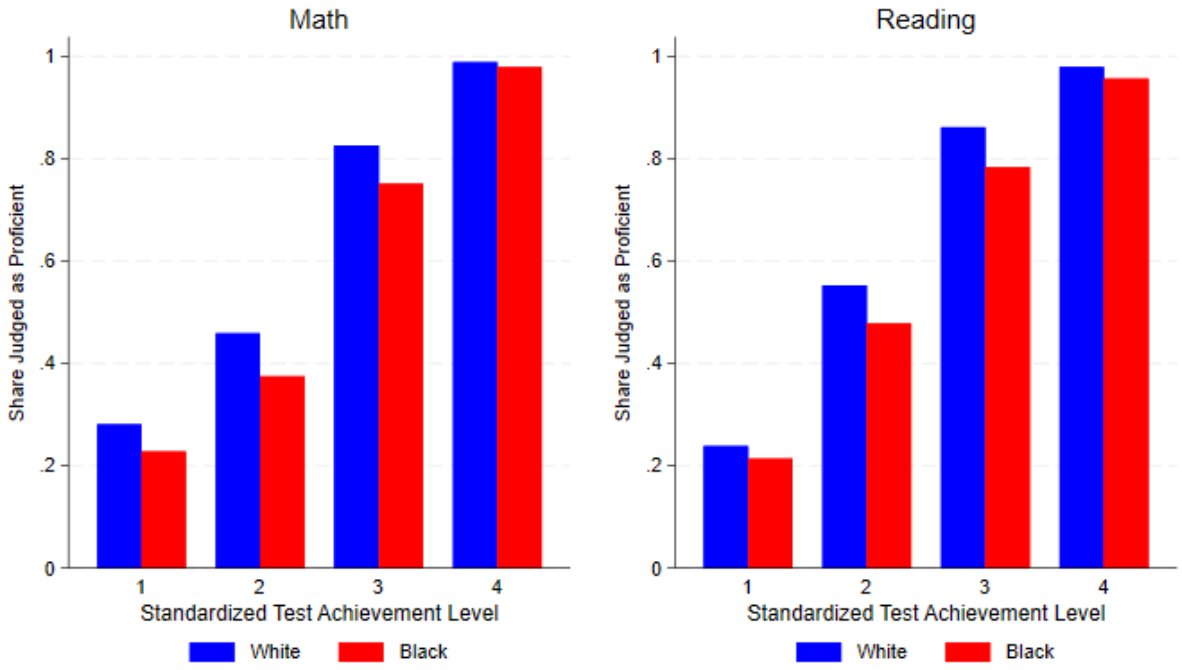


Note: Test scores are standardized within year and grade to have an aggregate mean of zero and standard deviation of one.

y-axis of the figures measures the share of students who are rated by their teachers as being proficient (i.e., at a level 3 or level 4) by their teachers in a subject. For each achievement level in math and reading, the share of White students scoring at that achievement level who are judged as proficient by teachers is larger than the share of Black students. These differences are significant at a 99 percent confidence level for each achievement level in both subjects.

Figure 2 presents suggestive evidence that teachers may display negative bias towards Black students relative to White counterparts, measured as teachers perceiving Black students as lower achieving than White peers with the same test-based achievement level. However, these graphs should not be interpreted as causal evidence of teacher bias for a several reasons. First, the breakdown of raw test scores into coarse bins may not fully capture underlying differences in test performance between Black and White students. It is plausible that some of the observed differences in teacher assessments between races within an achievement

Figure 2: Black-White Gap in Teacher Assessments across Standardized Test Scores



Note: Figures display the average share of students who are judge as have a proficiency of level 3 or level 4 by their teachers across discretized standardized test achievement level.

level reflect these underlying disparities rather than teacher bias. Second, the comparisons do not address the possibility of non-random sorting of students by race into different districts, schools, teachers, and/or classrooms, all of which could influence assessment scores. While racial evaluation gaps stemming from factors such as school-level differences in evaluation standards may still reflect systemic bias, they would not necessarily be attributable to biases at the teacher level. Lastly, these comparisons do not correct for measurement error in standardized scores, potentially resulting in misleading conclusions regarding racial differences in teacher assessment, even if students were randomly assigned to classrooms. In the subsequent sections, I discuss the approach used to identify the causal role of race on teacher evaluations of student achievement.

III Empirical Strategy

To assess racial differences in teacher assessments of students, I estimate the following linear probability model:

$$Y_{ic} = \mathbf{Race}_i'\beta + f(Test_{ic}) + \mathbf{X}_i'\Gamma + \zeta_c + \epsilon_{ic} \tag{1}$$

where Y_{ic} is an indicator variable for student i in class c that takes a value of one if the student's teacher assesses them as being proficient in the subject (i.e., if the teacher assesses a student as having an achievement level of 3 or 4), and zero otherwise. To capture racial differences across teacher assessments, I include a vector of student race indicator variables, $\mathbf{Race}_i'$. I control for underlying standardized test scores, $Test_{ic}$, to ensure racial differences in assessment are not being driven by achievement differences by race.[7] The vector $\mathbf{X}_{ic}'$ controls for a set of observable student characteristics, including gender and economic disadvantage status. These variables address the possibility that student composition differs across race along these characteristics, which could subsequently affect teacher assessments. Finally, I include a class fixed effect, ζ_c, which means β captures racial differences in outcomes within classrooms. Class fixed effects are, by definition, school-, teacher-, course-, and year-specific, thereby accounting for the possibility that non-random sorting of students along these dimensions is driving racial differences in outcomes. For instance, this addresses concerns such as some teachers being more lenient in assessments than others, classroom-specific shocks affecting learning, and changes in standardized testing standards or teacher assessment standards over time.

The coefficient vector of interest, β, measures the effect of student race on teachers' assessments, controlling for test achievement. The central identifying assumption underpinning this specification is that, conditional on students' standardized test scores, there are no unobserved factors correlated with race that influence racial disparities in teacher assessments

[7] I control for current year test scores rather than lagged test scores because current year test scores capture a student's achievement at the time of teacher assessment (since both occur at the end of the year). Teacher biases may influence achievement gain over the year, meaning differences in teachers' assessments by race using lagged test score controls could in part reflect real differences in achievement. Depending on the mechanism through which teacher bias manifests (e.g., biases leading to more negative assessments of Black students versus biases leading to lower expectations of Black students), this could lead to under- or over-estimates of the true effect of student race.

of achievement within a classroom. While this empirical strategy mitigates many external factors that could potentially influence teacher assessments, I discuss potential identification concerns and provide robustness checks for the validity of these assumptions in Section V.B.

IV Measurement Error

While the inclusion of standardized test score controls in Equation 1 is helpful in reducing the concern that estimation results are picking up underlying differences in achievement by race, test measurement error will also bias estimates of β. Prior studies have shown that standardized test scores measure student achievement with a significant amount of error (Boyd et al., 2013; Kane and Staiger, 2002). In this setting, test scores are subject to various sources of measurement error. First, there is inherent measurement error linked to the test instrument itself. Each test comprises a finite set of questions, leading to randomness in question selection and alignment with students' subject proficiency. Second, measurement error stems from variations in test conditions, such as student health or room temperature on exam day. Finally, as standardized tests are multiple-choice, a third source of measurement error arises from luck when in guessing answers to unfamiliar questions.

This section first illustrates econometrically how classical measurement error in standardized test scores will bias estimates of the relationship between student race and teacher evaluations. Next, I provide an overview of the extent of measurement error in standardized tests in North Carolina public schools. Following this, I discuss strategies for correcting for measurement error using instrumental variables. Finally, I perform Monte Carlo simulations to quantify the bias generated by measurement error in this setting and assess the abillity of instrumental variables to correct or bias in this setting.

IV.A Conceptual Framework

Consider the following equation capturing the relationship between teacher assessments (Y), standardized test scores (T), and student race (R):

$$Y = \beta R + \gamma T + \epsilon \tag{2}$$

Suppose that race does not affect teachers' assessments of students. Furthermore, instead of observing T, the econometrician observes a noisy measure of T that contains random error:

$$\tilde{T} = T + u \tag{3}$$

where the measurement error on T has mean zero ($E(u) = 0$) and is uncorrelated with T, R, and the equation error ($\text{plim}\frac{1}{n}(T'u) = 0$, $\text{plim}\frac{1}{n}(R'u) = 0$, $\text{plim}\frac{1}{n}(\epsilon'u) = 0$). A well known result from the econometric literature (Wooldridge (2010); Cameron and Trivedi (2005)) is that the OLS estimator for $\hat{\gamma}$ in this situation is:

$$\text{plim}\hat{\gamma} = \frac{\beta\left[\sigma_R^2\sigma_T^2 - (\sigma_{_{TR}})^2\right]}{\sigma_R^2(\sigma_T^2 + \sigma_u^2) - (\sigma_{_{TR}})^2} = \gamma\lambda \tag{4}$$

where $\lambda \leq 1$, so $\hat{\gamma}$ is biased towards zero. Furthermore, if R and T are positively correlated ($\rho_{_{TR}} > 0$), then the bias in $\hat{\beta}$ will go in the opposite direction of the bias in $\hat{\gamma}$ (DeGroot and Schervish, 2011). Intuitively, this is because measurement error in T reduces the explanatory power of T and biases $\hat{\gamma}$ towards zero. In turn, some of the true variation explained by T will be attributed to R instead. Thus, $\hat{\beta}$ will be positively biased in the case where $\rho_{_{TR}} > 0$. Conversely, it follows that if $\rho_{_{TR}} < 0$, the bias in $\hat{\beta}$ will go in the same direction of the bias in $\hat{\gamma}$ and thus be negatively biased.

This bias in $\hat{\beta}$ poses a significant issue in the setting of this paper, given that Figure 1 shows that standardized test scores are correlated with race. Specifically, since Black students have lower standardized test scores on average compared to White students, random measurement error will downwardly bias the coefficient estimate of teacher assessments for Black students in Equation 1. Thus, measurement error will lead coefficient estimates to suggest teachers are more negative in their assessments of Black students even in a setting where teachers do not differ in their assessments between a White and Black student who have the same underlying achievement.

IV.B Measurement Error in North Carolina

Holdzkom et al. (2010) provides a report of measurement error in standardized tests by

grade and subject in North Carolina for the 2010 school year. Their report provides values for the standard error of measurement (SEM) for raw test scores, which corresponds to σ_u in Equation 3. These values are displayed in column 2 of Table 2, and I report corresponding raw test score standard deviations from the data in column 3. Column 4 divides column 2 values by column 3 values to report how many test score standard deviations the raw test score SEM values represent. I find that the standard error of measurement in test scores in 2010 ranges from 0.282 to 0.513 standard deviations across grades and subjects. The mean value for the standard error of measurement is 0.455 standard deviations in math and 0.353 standard deviations in reading.

Table 2: Measurement Error in North Carolina Standardized Tests, 2010

Subject+grade	Raw test score SEM (σ_u) (2)	Raw test score SD (σ_T) (3)	Raw $\frac{\sigma_u}{\sigma_t}$ (4)
Math 3rd grade	4	8.92	0.448
Math 4th grade	3	8.61	0.348
Math 5th grade	4	8.39	0.477
Math 6th grade	4	8.46	0.473
Math 7th grade	4	8.53	0.469
Math 8th grade	4	7.79	0.513
Reading 3rd grade	3	10.62	0.282
Reading 4th grade	3	8.94	0.336
Reading 5th grade	3	8.14	0.369
Reading 6th grade	3	8.07	0.372
Reading 7th grade	3	8.06	0.372
Reading 8th grade	3	7.75	0.387

Overall, Table 2 illustrates that there is a sizable amount of measurement error in standardized test scores in North Carolina during the time period of my sample.[8] This table provides a benchmark for the magnitude of measurement error, which is used in simulation exercises in the Section to evaluate the expected bias due to measurement error in this context.

[8]This information has not been made public for other years in the sample to my knowledge, although Holdzkom et al. (2010) notes that in regard to measurement error, "historically it has been around 3–4 points."

IV.C Correcting For Measurement Error

A well-established method for correcting for measurement error in explanatory variables is to use an instrumental variables approach (Hausman, 2001; Bound et al., 2001). In this context, a valid instrument must meet two key criteria. First, it should be correlated with standardized test scores, and second, it should only influence teacher assessments through its impact on standardized test scores. This implies that while the instrument can also be measured with error, the error in the instrument should not be correlated with the error in the mis-measured explanatory variable. Conceptually, an instrument meeting these conditions can correct for measurement bias by isolating the portion of a student's standardized test score that reflects their achievement from the portion of the test score reflecting noise.

Prior research has employed various instruments to correct for potentially mis-measured standardized test scores. Botelho et al. (2015) and Bond and Lang (2017) instrument for test-based achievement using lagged achievement in the same subject. Zabel (2008) and Hanushek and Rivkin (2009) correct for measurement error in test scores for a given subject using a student's contemporaneous test score in a different subject.

Building off of prior studies, I correct for measurement error using various instruments. First, I instrument for standardized test scores in a subject using test scores in the other subject (i.e., instrumenting for math scores using reading scores and vice versa). One potential limitation of this instrument is that errors across contemporaneous subjects may be correlated if, for example, a student is unwell during the testing window. To address this, I also try instrumenting for standardized test scores using lagged test scores in the same subject. Although teachers are instructed to assess students on their current achievement, one potential limitation of this approach is that the validity of the instrument may not hold if teachers' assessments are colored by their impressions of students at the beginning of the year. This is because achievement in a subject at the beginning of the year is likely correlated with achievement in the subject at the end of the prior school year. To address this concern, I instrument for standardized test scores using lagged standardized test scores in the other subject, which is arguably the instrument least likely to suffer from validity concerns. I also use twice lagged same-subject and other-subject test scores to further reduce the likelihood

of correlated errors. However, incorporating additional lagged test score measures comes at the expense of sample size since I only observe test scores for students in grades 3-8.

IV.D Monte Carlo Simulations

Next, I use Monte Carlo simulations to demonstrate the effect of measurement error on parameter estimates in the setting of this study. The data-generating process closely resembles the model outlined in Equation 2, extended to have a more flexible approach to modeling standardized test scores:

$$Y = \beta R + \gamma_1 T + \gamma_2 T^2 + \gamma_3 T^3 + \gamma_3 T^4 + \epsilon \tag{5}$$

where students are either Black (B) or White (W), and R is an indicator variable such that $R = 1$ if $B = 1$ and $R = 0$ if $W = 1$. Standardized test scores are normalized to have a mean of zero and standard deviation of one. Underlying achievement A varies by race such that $A^W \sim \mathcal{N}(0.28, .95)$ and $A^B \sim \mathcal{N}(-0.50, 0.88)$. Underlying distributions reflect actual means and standard deviations in math test scores for my sample.[9] I assume that in the absence of measurement error, test assessments T are accurate and precise measures of achievement.

The top panel of Table 3 displays simulation results for $\hat{\beta}$ in the presence of classical measurement error. In the simulations, teachers do not display racial bias in their assessments of students, and teacher assessment, Y, is purely a function of ability, A. Specifically, Y is a binary variable that takes a value of one if $A > -0.643$ and zero otherwise. This threshold was chosen to match the actual share of students in the data who have a standardized test score corresponding to an achievement level of 3 or 4. To obtain β in this setting, I estimate Equation 5 using true achievement, A, instead of a noisy test score T. Results indicate $\beta = -0.0004$, indicating controlling for achievement using a fourth order polynomial is a reasonable specification choice.

To assess measurement error bias, I allow T to be measured with varying degrees of error and display estimated $\hat{\beta}$ values. The first row of simulation results in the top panel uses

[9]Simulation results using reading test score means and standard deviations can be found in Appendix B.A.

$\sigma_u = 0.455$, matching the mean level of measurement error reported in math standardized test scores by Holdzkom et al. (2010), as described in Section IV.B. Results show that the presence of this error leads to estimates of $\hat{\beta} = -0.057$, indicating teachers are 5.7 percentage points less likely to assess Black students as having proficient mastery of math compared to White peers with the same test scores, even though teachers do not demonstrate racial differences in their assessment standards. The next two rows indicate that measurement error is still sizable with lower values of σ_u, although lower σ_u leads to lower bias in $\hat{\beta}$, as expected. Results indicate that there would be a sizable amount of bias even with less than half of the reported levels of measurement error in the North Carolina data.

Table 3: Measurement error simulation results

Measurement error source	IV	β	$\hat{\beta}$	Standard error
Ordinary Least Squares				
$\tilde{T} = A + \mathcal{N}(0, 0.455)$		-0.0004	-0.0573	0.0022
$\tilde{T} = A + \mathcal{N}(0, 0.3)$		-0.0004	-0.0285	0.0020
$\tilde{T} = A + \mathcal{N}(0, 0.2)$		-0.0004	-0.0137	0.0019
Instrumental Variables				
$\tilde{T} = A + \mathcal{N}(0, 0.455)$	$Z = A + \mathcal{N}(0, 0.455)$	-0.0004	-0.0004	0.0025
$\tilde{T} = A + \mathcal{N}(0, 0.3)$	$Z = A + \mathcal{N}(0, 0.3)$	-0.0004	-0.0004	0.0021
$\tilde{T} = A + \mathcal{N}(0, 0.2)$	$Z = A + \mathcal{N}(0, 0.2)$	-0.0004	-0.0003	0.0019

Results run using 1,000 simulations, and the column reporting $\hat{\beta}$ represents the average estimate over these simulations. The sample size is $n = 100,000$ with $R = 0$ for 64% of observations and $R = 1$ for 34% of observations to match the ratio of White and Black students in my sample. Appendix B.B adds classroom fixed effects to the simulation, demonstrating that the magnitude of $\hat{\beta}$ is sensitive to the inclusion of these fixed effects, although the sign is not.

The bottom panel of able 3 presents simulation results using instrumental variables to correct for measurement error. I assume the existence of a valid instrument that is correlated with standardized test scores and only affects teacher assessments through its effect on standardized test scores. I allow the instrument to be measured with error as well, assuming noise in the instrument is uncorrelated with noise in standardized test scores. I simulate the estimated $\hat{\beta}$ values using this instrument. Simulation results indicate a valid instrument in this setting can be used to produce an unbiased estimator of β.

V Results

V.A Main Results

Table 4 presents OLS estimation results for Equation 1, providing estimates of the relationship between student race and teacher assessments, conditional on standardized test scores. Various specifications to control for standardized test scores are displayed across columns. In column 1, the preferred specification, test scores are included as a fourth order polynomial. Column 2 employs a third-order polynomial, while column 3 utilizes a linear control for test scores. In Column 4, raw test score dummy variables are interacted with year and grade. Test scores in Columns 1-3 are standardized within grade and year, ensuring a consistent interpretation of a one-unit change in test scores across grades and years.[10] Across specifications, the results are fairly consistent. Aligning with prior research in other contexts, the findings suggest that teachers evaluate Black students more unfavorably compared to their White counterparts within the same class and with identical standardized test scores. Specifically, in column 1, teachers are 0.4 percentage points less likely to rate Black students as proficient in math relative to White students with equivalent standardized test scores, representing a 0.5 percent reduction from the mean likelihood of proficiency assessment. Similarly, in reading, teachers are 2.0 percentage points less likely to assess Black students as proficient compared to their White peers with the same standardized test scores, corresponding to a 2.7 percent decrease from the mean likelihood of proficiency assessment.

Next, Table 5 presents results using instrumental variables to correct for measurement error in standardized test score assessments. Column 1 displays OLS estimation results using the preferred fourth order polynomial test score control from Table 4, showing that teachers are more negative in their assessments of Black students compared to White students in both math and reading, controlling for standardized test scores. Results using different instruments for standardized test scores are displayed across columns 2-6.

Column 2 uses lagged test scores in the tested subject to instrument for standardized test scores.[11] Estimation results indicate teachers are 2.3 percentage points more likely to assess

[10] Results remain robust to the use of raw test score controls instead of normalized ones.

[11] All IV estimations use fourth order test score instruments. Results are robust to alternative functional form specifications for test scores in both the first stage and the instrument, alleviating potential concerns

Table 4: OLS Estimates: Racial Differences in Teachers' Assessments of Students

$f(Test)$:	Fourth order polynomial (1)	Fourth order polynomial (2)	Linear (3)	Raw test score dummies (4)
Math				
Black	-0.004***	-0.005***	-0.010***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
N	2,403,904	2,403,904	2,403,904	2,403,879
Baseline mean	0.754	0.754	0.754	0.754
Reading				
Black	-0.020***	-0.020***	-0.022***	-0.020***
	(0.001)	(0.001)	(0.001)	(0.001)
N	2,396,569	2,396,569	2,396,569	2,396,526
Baseline mean	0.751	0.751	0.751	0.751

*** p<0.01, ** p<0.05, * p<0.1. SE clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include class fixed effects, controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Test scores in columns (1)-(3) are standardized to have a mean of zero and SD of one. Raw score dummies are year- and grade-specific. The sample is restricted to individuals who have information available for other subject test scores, lagged same- and other- subject test scores, and twice lagged same- and other-subject test scores in order to have a consistent sample across instrumental variables specifications. Results do not substantively change when the unrestricted sample is used.

Black students as proficient in math than White students with the same standardized test-based achievement and 0.9 percentage points more likely to in reading. Column 3 uses twice lagged test scores as an instrument. A potential advantage of twice lagged test scores over lagged test scores is that these scores are more removed from current test scores and thus less likely to suffer from correlated measurement errors. Estimations using this specification find teachers are 1.8 percentage points more likely to assess Black students as proficient in math than White students with the same standardized test-based achievement and 0.6 percentage points more likely to in reading. One potential concern with using lagged test scores in a given subject is that teachers' assessments of students may be influenced by students' academic achievement levels at the beginning of the year. While teachers are asked to assess

raised by Dieterle and Snell (2016).

students' proficiency at the time that they are surveyed, which occurs at the end of the year, it may still be the case that teachers' perceptions of achievement are influenced by achievement throughout the year. If this were the case, lagged test scores, which are likely correlated with student achievement at the beginning of the year, would not be a suitable instrument since they directly influence teachers' assessments of students.[12]

Table 5: IV Estimates: Racial Differences in Teachers' Assessments of Students

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Black	-0.004***	0.023***	0.018***	0.020***	0.027***	0.025***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,403,904	2,403,904	2,403,904	2,403,904	2,403,904	2,403,904
Baseline mean	0.754	0.754	0.754	0.754	0.754	0.754
First stage F-stat		1,755	1,035	755	491	326
Reading						
Black	-0.020***	0.009***	0.006***	0.028***	0.032***	0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,396,569	2,396,569	2,396,569	2,396,569	2,396,569	2,396,569
Baseline mean	0.751	0.751	0.751	0.751	0.751	0.751
First stage F-stat		1,743	1,276	738	670	505

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Sample is restricted to individuals who have information available for other subject test scores, lagged same- and other- subject test scores, and twice lagged same- and other-subject test scores in order to have a consistent sample across specifications.

In response to these concerns, column 4 uses an alternative instrument, standardized test scores in the other subject (math test scores for reading test scores and vice versa) to correct fore measurement error. Estimates using this instrument find that teachers are 2.0 percentage points more likely to assess Black students as proficient in math than they are White students with the same standardized test-based achievement. They are 2.8 percentage points more likely to assess Black students as proficient in reading. An advantage of this approach is

[12]I do not include lagged test score controls in my estimations since teacher bias may also influence student achievement gains over the year. Twice lagged test scores do not eliminate this issue since twice lagged test scores are highly correlated with lagged test scores.

that it circumvents the potential concern of using lagged test scores as an instrument by using achievement scores in a different domain. One concern with this approach is that measurement error in the instrument may be correlated with measurement error in the test score since the testing window typically takes place around the same time each year. For example, if a student was ill during the testing window, this could lead to mis-measurement in both math and reading scores in a given school year. To address this concern, column 5 uses lagged other subject test scores, which is perhaps the least likely of all the instruments thus far to suffer from endogeneity issues. Estimation results using this specification indicate teachers are more likely to assess Black students as proficient than they are White students with the same standardized test-based achievement by 2.7 percentage points in math and 3.2 percentage points in reading. Finally, column 6 uses twice lagged other subject test scores as an instrument. Results using this specification indicate teachers are more likely to assess Black students as proficient than they are White students with the same standardized test-based achievement by 2.5 percentage points in math and 2.8 percentage points in reading.[13]

Overall, results correcting for measurement error in standardized test scores paint a very different picture from OLS results. IV results indicate teachers are 1.8-2.7 percentage points more likely to evaluate Black students as grade-level proficient in math achievement than they are White peers with the same standardized test-based achievement, a 2.4-3.6 percent increase from the mean. Similarly, teachers are 0.6 to 3.2 percentage points more likely to evaluate Black students as proficient in reading than they are White peers with the same standardized test-based achievement in math, a 0.8-4.3 percent increase from the mean.

V.B Robustness Checks

The first robustness check I conduct assesses the concern that racial differences in teachers' assessments of students arise from differences in unobserved behavioral characteristics between students that are correlated with race. Despite explicit instructions for teachers to assess students solely based on academic proficiency, external factors could nevertheless in-

[13]The magnitude of these findings are considerably smaller than magnitudes found in the simulation exercises in Table 2, which can be explained by the simulation exercises not containing classroom fixed effects. Appendix B.B shows that bias is sensitive to the degree of correlation of student ability within classes.

fluence their perceptions. To address this concern, I re-estimate results using a specification that includes controls for lagged school attendance and lagged disciplinary incident history in Table A1 in Appendix A.[14] Results using this specification are very similar to those of the main specification in Table 5: I find that teachers are less likely to assess Black students as proficient compared to White students with the same standardized test scores using an OLS estimation. However, after correcting for measurement error, I consistently find that teachers are *more* likely to assess Black students as being proficient in math and reading than White students, conditional on standardized test-based achievement. Overall, controlling for behavioral characteristics leads to similar findings as the main results, suggesting my results are not driven by unobserved correlates of race.

Next, I address the key identifying assumption that the instrument does not directly influence teachers' assessments of students. This assumption may be invalid if teachers have knowledge regarding a student's performance in the instrumented area (e.g., how a student performed in a prior year and/or in a different subject), and this information influences their assessment of a student's proficiency in the current subject. While it is not possible to prove whether an instrument satisfies this requirement, I conduct a robustness check by re-estimating my specification using a subset of observations where the current teacher differs from the teacher for the instrumented course. For instance, in specifications where the instrument is a student's test score in another subject, I limit the sample to observations where the student's math and reading teachers are different. Similarly, in specifications where the instrument is the student's lagged test score in the subject, I restrict the sample to observations where the student had a different teacher in the subject last year. The rationale behind this robustness check is that teachers who previously taught a student in a different course are more likely to have strong perceptions of the student's performance in that course, potentially allowing that information to influence their assessment in the course of interest. Results using this restricted sample, presented in Table A2 in Appendix A, are

[14] I include lagged behavioral controls instead of current year controls because student behavior may be endogenous to teacher biases and subsequent treatment of students. Even lagged behavioral controls may be problematic in adjusting for unobserved behavioral differences by race in this setting since prior research has found that teachers display significant racial biases in adjudicating disciplinary incidents, which is why the main specification does not include these controls (Barrett et al., 2021; Shi and Zhu, 2022; Gershenson et al., 2022).

reassuringly consistent with those from the full sample.

As a second robustness check to the concern that the instruments used may directly influence teachers' assessments of students, I use future test scores as alternative instruments. The advantage of using a future instrument is that future test scores cannot influence teachers' assessments of students in the current year, which makes them an attractive instrument for addressing this concern. However, a potential drawback of this instrument is that teachers' perceptions of students may directly influence their future test scores, independently of effects on current test scores.[15] For this robustness check, I instrument for standardized test scores using standardized test scores in the same subject in the following year and standardized test scores in the other subject in the following year. Results of this analysis are displayed in Table A3 in Appendix A, and estimates are qualitatively similar to those of the main results.

Relatedly, teachers may be influenced by students' entry-level achievement. Despite being instructed to evaluate students' achievement at the end of the year, teachers may form perceptions throughout the academic term. In contrast, standardized tests assess students' knowledge at the end of the year. Consequently, if two students achieve the same standardized test scores at the end of the year, a teacher might assess them differently if one started the year with lower achievement. This raises two identification concerns. First, estimates of the relationship between student race and teacher assessment could suffer from omitted variables bias if achievement growth during the year correlates with race. Second, instruments relying on prior year test scores in the same subject (e.g., lagged test scores and twice lagged test scores) may not solely impact the outcome through their relationship with the endogenous variable. To mitigate this, Table A4 in Appendix A presents a robustness check that incorporates lagged test scores, in addition to current test scores, to control for entry-level achievement. Column 1 displays OLS results, while column 2 shows IV results correcting for measurement error. I instrument for test scores and lagged test scores using scores from other subjects and lagged scores from other subjects. In line with the main results, findings in column 2 indicate that teachers are 2.8 percentage points more likely to

[15]For example, if teacher perceptions of students have effects on long term growth, this may not be fully captured through effects on contemporaneous test scores.

assess Black students as proficient in math and 3.1 percentage points more likely to assess Black students as proficient in reading compared to White students with the same current and lagged test score achievement.

Finally, I address concerns regarding the identification of effects from a limited support. Specifically, the main specification incorporates class fixed effects, meaning racial disparities in teacher assessments are identified from classes with both White and Black students. The majority of classes in my sample–71.2–contain students of both races, significantly alleviating this concern. As an additional robustness check, Tables A5 and A6 in Appendix A replace classroom fixed effects with teacher and school fixed effects, respectively. Using these coarser fixed effects offers the advantage that the vast majority of teachers and schools observing both White and Black students: 96.4% of teachers and 99.8% of schools have students of both races during the sample period. However, the drawback of employing these fixed effects is that they may not fully address sorting that correlates with teachers' assessments. For example, it is possible that a teacher instructs both algebra and pre-algebra math classes to students in the same grade and tends to assign higher assessments to students in the algebra class based on the assumption that they are more proficient due to their enrollment in algebra. If students non-randomly sort into algebra or pre-algebra across race, this could result in course assessment disparities being attributed to racial differences. Results in Tables A5 and A6 in Appendix A suggest that findings do not substantially change when substituting classroom fixed effects with teacher or school fixed effects.

V.C Heterogeneity and Mechanism Exploration

There are various mechanisms that could contribute to Black students receiving higher assessments than White students relative to their standardized test scores. Broadly, these results may stem from either racial assessment disparities in the testing instrument or from teachers. First, racial gaps in assessments might reflect biases in standardized tests, particularly if the tests exhibit negative biases against Black students. These findings align with previous research highlighting test score biases against students from marginalized groups (Sandoval, 1979; Freedle, 2003; Duncan and Sandy, 2013). Under this mechanism, it is also possible that teachers are negatively biased in assessments of Black students, if standardized

test bias effects dominate teacher bias effects.

Alternatively, the results may stem from biases exhibited by teachers, leading to differential assessment standards for Black students compared to White students. One possible channel is that teachers adjust their assessments standards to account for student background. For example, teachers may perceive that a student from a more disadvantaged background who achieves the same academic mastery as a student from a more privileged background demonstrates more impressive achievement and adjust their assessments to reflect this. Second, it could be that teacher biases manifest in the form of teachers having lower expectations for Black students relative to White students, occurring in response to negative biases or stereotypes towards Black students. If teachers have lower expectations for a group, this could then imply that the bar for achievement is lower for members of that group. Thus, for Black and White students with the same level of achievement, teachers may evaluate Black students as being higher achieving. Third, results could be driven by social desirability bias if teachers inflate their assessments of Black students to conform to responses they feel are more socially favorable. This mechanism is also compatible with negative teacher bias that manifests as lower conditional assessments of Black students. This mechanism may also coexist with negative teacher biases that depress assessments of Black students if social desirability bias effects dominate teacher bias effects. I refer to these mechanisms as:

- M1: Standardized test bias against Black students (with potential negative teacher bias)
- M2: Teachers account for student background in their assessments
- M3: Teachers haver lower expectations of achievement for Black students, leading to lower assessment standards
- M4: Teachers inflate assessments of Black students due to social desirability bias (with potential negative teacher bias)

I explore these mechanisms in this section through additional analyses. I first examine whether racial disparities in teacher assessments vary depending on the race of the teacher.

Prior research has found that Black teachers have higher expectations of Black students relative to White students compared to White teachers (Papageorge et al., 2020). If Mechanism 3, positing that teachers have lower expectations for Black students, is the driving force behind the results, I would anticipate smaller racial differences in assessments for Black teachers compared to White teachers. If bias originates from standardized tests, teacher characteristics may not influence the size of racial assessment gaps. However, given that standardized test bias can coexist with negative teacher bias, it's plausible that racial assessment gaps may be larger for Black teachers as well.[16] Predictions regarding differences in racial assessment gaps by teacher race are less clear for the other mechanisms.

Table 6 displays the analysis findings regarding racial differences in teachers' assessments of students by teacher race. The analysis is confined to classes instructed by White or Black teachers, constituting 81% of the teachers in the sample. The results reveal no discernible differences in the racial assessment gap between Black and White teachers in either math or reading after adjusting for measurement error. These findings contradict the predictions outlined in Mechanism 3.[17]

Next, I examine differences in teachers' assessments of students based on economic disadvantage status, while controlling for achievement and race. Students are classified as economically disadvantaged if they qualify for free or reduced-price lunch. Under Mechanism 2, I expect teachers to assess economically disadvantaged students more positively than their non-economically disadvantaged peers since teachers adjust their assessments based on student background. Predictions regarding differences in racial assessment gaps by teacher race are less straightforward for the remaining mechanisms, which are race-specific and thus do not necessarily have direct implications for assessment gaps by socioeconomic status. Table 7 presents the analysis results. Contrary to the predictions of Mechanism 2, I find that

[16]Chin et al. (2020) finds that Black teachers display lower implicit bias towards Black students compared to White teachers, suggesting they would be more positive in their assessments of Black students.

[17]These findings differ from those in Ouazad (2014), which finds that teachers assess same-race students more positively than they do different-race students, conditional on test score performance. The paper finds that an implausibly large amount of measurement error would need to be present to cancel out these results. The data used in that study come from the Early Childhood Longitudinal Study, Kindergarten Class (ECLS-K) of 19981999, and some differences between that study and this one include the population covered, time period, and grade span, all of which may contribute to differences in findings. Additionally, teacher assessments on the ECLS-K consist of up to 20 items that measure various skills, rather than one comprehensive measure of achievement.

Table 6: Heterogeneity by Teacher Race: Racial Differences in Teachers' Assessments of Students

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Black	-0.009***	-0.009***	-0.009***	-0.009***	-0.010***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black×Black teacher	0.006**	0.000	0.000	0.001	-0.002	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	2,329,448	2,329,448	2,329,448	2,329,448	2,329,448	2,329,448
Reading						
Black	-0.021***	0.009***	0.006***	0.027***	0.032***	0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black×Black teacher	0.007***	0.001	0.002	0.003	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	2,347,119	2,347,119	2,347,119	2,347,119	2,347,119	2,347,119

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Teacher race is interacted with all controls and test scores. Sample is restricted to individuals who have information available for other subject test scores, lagged same- and other- subject test scores, and twice lagged same- and other-subject test scores in order to have a consistent sample across specifications. Sample is also restricted to classes taught by White or Black teachers.

both in math and reading, teachers are less inclined to assess economically disadvantaged students as proficient compared to peers of the same race/ethnicity who are not economically disadvantaged.

Finally, I examine whether racial differences in assessments vary based on the Black-White socioeconomic gaps within communities. Under Mechanism 2, I anticipate that teachers would adjust assessments more in areas with larger socioeconomic disparities between White and Black families, as Black students face comparatively greater socioeconomic barriers in such settings. Similarly, under Mechanism 3, I would anticipate expectations for Black students to be relatively lower compared to White students in areas with larger White-Black socioeconomic disparities. Consequently, this might manifest as relatively higher assessments for Black students in these areas. Predictions regarding differences in racial assessment gaps by teacher race are less clear for the remaining mechanisms.

Table 7: Socioeconomic Differences in Teachers' Assessments of Students

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Economically disadvantaged	-0.038***	-0.019***	-0.023***	-0.021***	-0.017***	-0.018***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,403,904	2,403,904	2,403,904	2,403,904	2,403,904	2,403,904
Reading						
Economically disadvantaged	-0.049***	-0.028***	-0.030***	-0.014***	-0.011***	-0.014***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,396,569	2,396,569	2,396,569	2,396,569	2,396,569	2,396,569

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Sample is restricted to individuals who have information available for other subject test scores, lagged same- and other- subject test scores, and twice lagged same- and other-subject test scores in order to have a consistent sample across specifications.

To address this question, I assess racial gaps in family income using data from the American Community Survey (ACS) from 2007-2013. The ACS allows for the breakdown of income gaps at the county level, which is well-suited for this analysis given that school districts in North Carolina are delineated at the county level. One limitation of this approach is that the ACS does not provide county identifier information for less populated areas in the state, restricting the analysis to 25 out of the 100 counties. However, this still encompasses 62% of observations in my sample, as the most densely populated counties are included.

I restrict my analysis in the ACS to individuals who have children living at home to align more closely with the population of families in the NCERDC. Subsequently, I compute the disparity in mean family income between White families and Black families for each county. Table A7 presents the analyses on how racial differences in teachers' assessments of students vary based on the White-Black income gap within the county where the student attends school. Results indicate teachers are relatively more negative in their assessments of Black students in both math and reading in counties where the White-Black income gap is higher.[18] These findings do not align with the predictions of Mechanisms 2 and 3.

[18]As a robustness check, I measure socioeconomic status as the White-Black difference in shares of individuals with bachelor's degrees for individuals with children in the house and find qualitatively similar results.

Table 8: Heterogeneity by County Racial Income Differences: Racial Differences in Teachers' Assessments of Students

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Black	-0.007***	0.020***	0.015***	0.018***	0.024***	0.023***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black×White-Black Income Gap (z-score)	-0.006***	-0.007***	-0.007***	-0.008***	-0.008***	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	1,464,352	1,464,352	1,464,352	1,464,352	1,464,352	1,464,352
Reading						
Black	-0.024***	0.005***	0.002*	0.023***	0.028***	0.024***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black×White-Black Income Gap (z-score)	-0.009***	-0.009***	-0.009***	-0.009***	-0.010***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	1,460,740	1,460,740	14,60,740	1,460,740	1,460,740	1,460,740

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, controls for Asian, Hispanic, other race, gender, and economic disadvantage status, and full interaction terms between the White-Black income gap and these variables. Sample is restricted to individuals who have information available for other subject test scores, lagged same- and other- subject test scores, and twice lagged same- and other-subject test scores in order to have a consistent sample across specifications. Sample is also restricted to the 25 counties that are identifiable in the American Community Survey.

Table 9 presents predictions of each mechanism in relation to the analyses conducted above, alongside the actual findings. Overall, the results in this section run counter to the predictions of Mechanisms 2 and 3. However, while these findings provide suggestive evidence, it is important to note that definitively pinpointing mechanisms requires additional data. For instance, datasets containing specific standardized test question items could facilitate item response theory analysis to more accurately identify sources of test bias. Alternatively, information on implicit bias association test results for individual teachers could offer insights into how teachers' biases influence their assessments.

Table 9: Mechanism Predictions

Test:	Black teacher	Economically disadvantaged	White-Black income gap
Mechanism predictions			
M1: Test bias	(+)	(+/−)	(+/−)
M2: Teacher accounts for background	(+/−)	+	+
M3: Teacher lowers expectations	−	(+/−)	+
M4: Teacher social desirability bias	(+/−)	(+/−)	(−)
Actual	( )	−	+/−

Symbols denote the predicted sign of the coefficient of interest. Parentheses denote results are compatible with null effects, and (+/−) denotes ambiguous sign predictions.

Finally, in Appendix A, I present additional heterogeneity analyses that are not directly linked to mechanism exploration. Table A8 examines results separately for elementary and middle school students to discern if the findings are driven by a specific age group. The analysis reveals that teachers are more positive in their assessments of Black students compared to White students in both elementary and middle schools. Next, Table A9 explores heterogeneity by teacher age and experience. Since there is a high degree of correlation between teacher age and experience, I jointly assess their roles. I categorize teachers as having low experience if they have four or fewer years of teaching experience, medium experience if they have 5-14 years of experience, and high experience if they have more than 15 years. Following the classification by Nguyen et al. (2023), I define teachers as being in the Baby Boomer generation if they were born between 1946 and 1964, the Gen X generation if they were born between 1965 and 1980, and the Millennial generation if they were born after 1980. The results indicate that in both math and reading, Millennial teachers assess Black students more negatively than Baby Boomer teachers, while there are no significant differences between the assessments of Gen X teachers and Baby Boomer teachers. Additionally, teachers with medium and high levels of experience tend to be more negative in their assessments of students compared to more novice teachers.

VI Conclusion

This study provides new insights on the nature of racial gaps in teachers' assessments of student achievement. I compare teachers' assessments of Black and White students in the same class, controlling for students' standardized test-based measures of achievement. After correcting for measurement error, I find that teachers assess Black students as being higher performing than White peers in the same class with the same standardized test-based achievement. Further analyses indicate these effects are consistent with a couple of different mechanisms. First, these results could be driven by standardized test bias against Black students relative to White students. Second, results are consistent with a setting in which social desirability bias drives teachers to inflate their assessments of Black students.

The findings in this paper carry several important implications. First, results from this

study emphasize the importance of addressing measurement error in empirical research, including settings in which the mis-measured regressor is not the variable of interest. In particular, a large number of studies currently in education currently use test scores or lagged test scores as a control variable. This study emphasizes the importance of correcting for test score measurement error when the regressor of interest is correlated with test scores. Second, this study provides novel evidence that at least in some settings, teachers are more positive in their assessments of students from disadvantaged racial groups. Third, findings raise caution against relying solely on standardized test scores as an objective measure of student achievement when examining racial disparities in achievement, a common practice in many studies. Fourth, the findings in this paper emphasize the importance of further research with additional data to better understand sources of discrepancy between teacher assessments and standardized test assessments of student achievement.

Ultimately, the findings in this paper warrant further exploration, regardless of whether racial disparities originate from teachers or tests, as both are pivotal in shaping students' academic trajectories. Teachers and tests alike play integral roles in many schools, influencing decisions regarding remediation and recommendations to various academic tracks. Thus, understanding the mechanisms behind racial discrepancies in teacher assessments and standardized test scores is crucial for ensuring equitable educational opportunities for all students.

References

Barrett, N., McEachin, A., Mills, J. N., and Valant, J. (2021). Disparities and Discrimination in Student Discipline by Race and Family Income. *Journal of Human Resources*, 56(3):711–748.

Bond, T. and Lang, K. (2017). The Black-White Education Scaled Test-Score Gap in Grades K-7. *Journal of Human Resources*, 53:0916–8242R.

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

Bound, J., Brown, C., and Mathiowetz, N. (2001). Chapter 59 - Measurement Error in Survey Data. In Heckman, J. J. and Leamer, E., editors, *Handbook of Econometrics*, volume 5, pages 3705–3843. Elsevier.

Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J. (2013). Measuring Test Measurement Error: A General Approach. *Journal of Educational and Behavioral Statistics*, 38(6):629–663. Publisher: American Educational Research Association.

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

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge, unknown edition edition.

Chin, M. J., Quinn, D. M., Dhaliwal, T. K., and Lovison, V. S. (2020). Bias in the Air: A Nationwide Exploration of Teachers Implicit Racial Attitudes, Aggregate Bias, and Student Outcomes. *Educational Researcher*, 49(8):566–578. Publisher: American Educational Research Association.

DeGroot, M. H. and Schervish, M. J. (2011). *Probability and Statistics*. Pearson, Boston, 4th edition edition.

Dieterle, S. G. and Snell, A. (2016). A simple diagnostic to investigate instrument validity and heterogeneous effects when using a single instrument. *Labour Economics*, 42:76–86.

Duncan, K. and Sandy, J. (2013). Using the Blinder-Oaxaca Decomposition Method to Measure Racial Bias in Achievement Tests. *The Review of Black Political Economy*, 40(2):185–206. Publisher: SAGE Publications Inc.

Francis, D. V. (2012). Sugar and Spice and Everything Nice? Teacher Perceptions of Black Girls in the Classroom. *The Review of Black Political Economy*, 39(3):311–320. Publisher: SAGE Publications Inc.

Francis, D. V., Oliveira, A. C. M. d., and Dimmitt, C. (2019). Do School Counselors Exhibit Bias in Recommending Students for Advanced Coursework? *The B.E. Journal of Economic Analysis & Policy*, 19(4). Publisher: De Gruyter.

Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, 73(1):1–43. Place: US Publisher: Harvard Education Publishing Group.

Gershenson, S., Hart, C. M. D., Hyman, J., Lindsay, C. A., and Papageorge, N. W. (2022). The Long-Run Impacts of Same-Race Teachers. *American Economic Journal: Economic Policy*.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study. *Journal of Political Economy*, 127(4):1826–1863. Publisher: The University of Chicago Press.

Griliches, Z. (1977). Estimating the Returns to Schooling: Some Econometric Problems. *Econometrica*, 45(1):1–22. Publisher: [Wiley, Econometric Society].

Hanushek, E. and Rivkin, S. (2009). Harming the best: How schools affect the blackwhite achievement gap -. *Journal of Policy Analysis and Management*, 28(3):366–393.

Hausman, J. (2001). Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left. *Journal of Economic Perspectives*, 15(4):57–67.

Hill, A. J. and Jones, D. B. (2021). Self-Fulfilling Prophecies in the Classroom. *Journal of Human Capital*. Publisher: The University of Chicago PressChicago, IL.

Holdzkom, D., Sumner, B., and McMillen, B. (2010). A Brief Look at: Test Scores and the Standard Error of Measurement. E&R Report No. 10.13. Technical report, Wake County Public School System. Publication Title: Wake County Public School System ERIC Number: ED564348.

Kane, T. J. and Staiger, D. O. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives*, 16(4):91–114.

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

Lavy, V. and Megalokonomou, R. (2019). Persistency in Teachers Grading Bias and Effects on Longer-Term Outcomes: University Admissions Exams and Choice of Field of Study.

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

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

Modalsli, J. and Vosters, K. (2022). Spillover bias in multigenerational income regressions. *Journal of Human Resources*. Publisher: University of Wisconsin Press Section: Articles.

Nguyen, N., Ost, B., and Qureshi, J. (2023). OK Boomer: Generational Differences in Teacher Quality.

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

Papageorge, N. W., Gershenson, S., and Kang, K. M. (2020). Teacher Expectations Matter. *The Review of Economics and Statistics*, 102(2):234–251.

Rangel, M. and Shi, Y. (2021). First Impressions: The Case of Teacher Bias. *Working Paper*.

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

Sandoval, J. (1979). The WISC-R and internal evidence of test bias with minority groups. *Journal of Consulting and Clinical Psychology*, 47(5):919–927. Place: US Publisher: American Psychological Association.

Shi, Y. and Zhu, M. (2022). Equal time for equal crime? Racial bias in school discipline. *Economics of Education Review*, 88:102256.

Shi, Y. and Zhu, M. (2023). Model minorities in the classroom? Positive evaluation bias towards Asian students and its consequences. *Journal of Public Economics*, 220:104838.

van Huizen, T., Jacobs, M., and Oosterveen, M. (2024). Teacher Bias or Measurement Error? *Working Paper*.

Ward, Z. (2023). Intergenerational Mobility in American History: Accounting for Race and Measurement Error. *American Economic Review*, 113(12):3213–3248.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Zabel, J. E. (2008). The Impact of Peer Effects on Student Outcomes in New York City Public Schools. *Education Finance and Policy*, 3:197–249. Publisher: MIT Press.

A Additional Tables

Table A1: Robustness Check: Racial Differences in Teachers' Assessments of Students with Behavioral Controls

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Black	-0.009***	0.019***	0.015***	0.017***	0.024***	0.022***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Lagged Absences	Y	Y	Y	Y	Y	Y
Lagged Disciplinary Incidents	Y	Y	Y	Y	Y	Y
N	2,402,132	2,402,132	2,402,132	2,402,132	2,402,132	2,402,132
Reading						
Black	-0.024***	0.004***	0.001*	0.021***	0.026***	0.022***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Lagged Absences	Y	Y	Y	Y	Y	Y
Lagged Disciplinary Incidents	Y	Y	Y	Y	Y	Y
N	2,394,755	2,394,755	2,394,755	2,394,755	2,394,755	2,394,755

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Sample is restricted to individuals who have information available for other subject test scores, lagged same- and other- subject test scores, and twice lagged same- and other-subject test scores in order to have a consistent sample across specifications.

Table A2: Robustness Check: Racial Differences in Teachers' Assessments of Students whose Teachers Differ from the Instrument Teacher

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Black	-0.004***	0.023***	0.019***	0.021***	0.027***	0.025***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,403,904	1,902,533	1,912,938	1,553,418	1,957,379	1,560,732
Reading						
Black	-0.020***	0.010***	0.007***	0.030***	0.034***	0.032***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,396,569	1,901,517	1,898,277	1,542,205	1,953,245	1,554,915

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status.

Table A3: Robustness Check: Racial Differences in Teachers' Assessments of Students using Future Test Scores as Instruments

	Instrumental variables	
	Future Subject (1)	Future Other Sub. (2)
Math		
Black	0.020***	0.030***
N	3,232,448	3,232,448
Reading		
Black	0.003***	0.032***
	(0.001)	(0.001)
N	3,222,772	3,222,772

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status.

Table A4: Robustness Check: Racial Differences in Teachers' Assessments of Students with Lagged Test Score Controls

	OLS	IV
	(1)	(2)
Math		
Black	0.011***	0.028***
	(0.001)	(0.001)
Test Scores	Y	Y
Lagged Test Scores	Y	Y
N	2,403,904	2,403,904
Reading		
Black	-0.007***	0.031***
	(0.001)	(0.001)
Test Scores	Y	Y
Lagged Test Scores	Y	Y
N	2,396,569	2,396,569

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Column 2 instruments for test scores and lagged test scores using other subject test scores and lagged other subject test scores.

Table A5: Robustness Check: Racial Differences in Teachers' Assessments of Students using Teacher Fixed Effects

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Black	-0.006***	0.024***	0.019***	0.020***	0.027***	0.025***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,410,828	2,410,828	2,410,828	2,410,828	2,410,828	2,410,828
Reading						
Black	-0.020***	0.012***	0.009***	0.029***	0.033***	0.030***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,404,095	2,404,095	2,404,095	2,404,095	2,404,095	2,404,095

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Column 2 instruments for test scores and lagged test scores using other subject test scores and lagged other subject test scores.

Table A6: Robustness Check: Racial Differences in Teachers' Assessments of Students using School Fixed Effects

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Black	-0.009***	0.023***	0.018***	0.018***	0.026***	0.024***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,411,631	2,411,631	2,411,631	2,411,631	2,411,631	2,411,631
Reading						
Black	-0.020***	0.012***	0.009***	0.029***	0.033***	0.030***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	2,404,095	2,404,095	2,404,095	2,404,095	2,404,095	2,404,095

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Column 2 instruments for test scores and lagged test scores using other subject test scores and lagged other subject test scores.

Table A7: Heterogeneity by County Education Differences: Racial Differences in Teachers' Assessments of Students

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Black	-0.007***	0.021***	0.015***	0.018***	0.024***	0.023***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black×White-Black BA Gap (z-score)	-0.007***	-0.008***	-0.007***	-0.008***	-0.008***	-0.008***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	1,464,352	1,464,352	1,464,352	1,464,352	1,464,352	1,464,352
Reading						
Black	-0.024***	0.005***	0.002*	0.023***	0.028***	0.024***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black×White-Black BA Gap (z-score)	-0.009***	-0.009***	-0.009***	-0.009***	-0.010***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	1,460,740	1,460,740	14,60,740	1,460,740	1,460,740	1,460,740

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, controls for Asian, Hispanic, other race, gender, and economic disadvantage status, and full interaction terms between the White-Black bachelor's degree gap and these variables Sample is restricted to individuals who have information available for other subject test scores, lagged same- and other- subject test scores, and twice lagged same- and other-subject test scores in order to have a consistent sample across specifications. Sample is also restricted to the 25 counties that are identifiable in the American Community Survey.

Table A8: Heterogeneity by Grade: Racial Differences in Teachers' Assessments of Students

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math: Elementary School						
Black	-0.010***	0.019***	0.016***	0.028***	0.036***	0.036***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	613,882	613,882	613,882	613,882	613,882	613,882
Math: Middle School						
Black	-0.002**	0.024***	0.019***	0.018***	0.023***	0.021***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	1,789,943	1,789,943	1,789,943	1,789,943	1,789,943	1,789,943
Reading: Elementary School						
Black	-0.024***	0.009***	0.007***	0.027***	0.032***	0.029***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	613,454	613,454	613,454	613,454	613,454	613,454
Reading: Middle School						
Black	-0.019***	0.009***	0.005***	0.028***	0.032***	0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	1,783,055	1,783,055	1,783,055	1,783,055	1,783,055	1,783,055

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Sample is restricted to individuals who have information available for other subject test scores, lagged same- and other- subject test scores, and twice lagged same- and other-subject test scores in order to have a consistent sample across specifications. Middle school includes grades 6-8, and elementary school (the omitted group) includes grades 3-5. (Note: in practice, elementary school students only include students in grades 4-5 since third graders will be dropped for not having lagged test scores.)

Table A9: Heterogeneity by Teacher Age and Experience: Racial Differences in Teachers' Assessments of Students

		Instrumental variables				
	OLS	Lagged Subject	Twice Lagged Subject	Other Subject	Lagged Other Sub.	Twice Lagged Other Sub.
	(1)	(2)	(3)	(4)	(5)	(6)
Math						
Black×Gen X	-0.002	-0.003	-0.003	-0.003	-0.003	-0.003
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Black×Millennial	-0.003	-0.006**	-0.006**	-0.006**	-0.006**	-0.006**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Black×Medium Experience	-0.004	-0.007***	-0.007***	-0.007***	-0.008***	-0.008***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Black×High Experience	-0.006**	-0.011***	-0.010***	-0.010***	-0.011***	-0.011***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Black	0.001	0.032***	0.027***	0.029***	0.036***	0.035***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
N	2,342,970	2,342,970	2,342,970	2,342,970	2,342,970	2,342,970
Reading						
Black×Gen X	-0.002	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Black×Millennial	-0.006**	-0.006**	-0.006**	-0.006**	-0.007**	-0.006**
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Black×Medium Experience	-0.008***	-0.010***	-0.010***	-0.011***	-0.011***	-0.011***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
Black×High Experience	-0.011***	-0.013***	-0.013***	-0.014***	-0.014***	-0.014***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Black	-0.011***	0.019***	0.016***	0.038***	0.043***	0.039***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
N	2,328,300	2,328,300	2,328,300	2,328,300	2,328,300	2,328,300

*** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the teacher level. The outcome is a binary variable taking a value of one if a teacher assesses a student's proficiency level at 3 or 4 and zero otherwise. The omitted racial/ethnic group is White students. All specifications include fourth order polynomial controls for standardized test scores (normalized to have a mean of zero and standard deviation of one) and classroom fixed effects, as well as controls for Asian, Hispanic, other race, gender, and economic disadvantage status. Sample is restricted to individuals who have information available for other subject test scores, lagged same- and other- subject test scores, and twice lagged same- and other-subject test scores in order to have a consistent sample across specifications. I categorize teachers as low experience if they have four or fewer years of teaching experience, medium experience if they have 5-14 years of experience, and high experience if they have more than 15 years. Following (Nguyen et al., 2023), I define teachers as being in the Baby Boomer generation if they were born 1946-1964, the Gen X generation if they were born 1965-1980, and the Millennial generation if they were born after 1980. The omitted experience group is low experience, and the omitted age group is Baby Boomers.

B Simulations Appendix

B.A Simulation results using reading distributions

Table B1 provides simulation results using underlying distributions of reading standardized test scores by race in reading. Specifically, I set $A^W \sim \mathcal{N}(0.30, .93)$ and $A^B \sim \mathcal{N}(-0.45, 0.91)$, reflecting actual means and standard deviations in reading test scores for my sample. I set $\sigma_u = 0.353$ to match the mean level of measurement error reported in reading standardized test scores by Holdzkom et al. (2010), as described in Section IV.B.

Table B1: Measurement error simulation results using reading simulations

Measurement error source	IV	β	$\hat{\beta}$	Standard error
Ordinary Least Squares				
$\tilde{T} = A + \mathcal{N}(0, 0.353)$		-0.0014	-0.418	0.0022
$\tilde{T} = A + \mathcal{N}(0, 0.3)$		-0.0014	-0.0317	0.0020
$\tilde{T} = A + \mathcal{N}(0, 0.2)$		-0.0014	-0.0157	0.0019
Instrumental Variables				
$\tilde{T} = A + \mathcal{N}(0, 0.353)$	$Z = A + \mathcal{N}(0, 0.353)$	-0.0014	-0.0015	0.0023
$\tilde{T} = A + \mathcal{N}(0, 0.3)$	$Z = A + \mathcal{N}(0, 0.3)$	-0.0014	-0.0015	0.0021
$\tilde{T} = A + \mathcal{N}(0, 0.2)$	$Z = A + \mathcal{N}(0, 0.2)$	-0.0014	-0.0015	0.0019

Results run using 1,000 simulations, and the column reporting $\hat{\beta}$ represents the average estimate over these simulations. The sample size is $n = 100,000$ with $R = 0$ for 64% of observations and $R = 1$ for 34% of observations to match the ratio of White and Black students in my sample. Appendix B.B adds classroom fixed effects to the simulation, demonstrating that the magnitude of $\hat{\beta}$ is sensitive to the inclusion of these fixed effects, although the sign is not.

B.B Simulation results with the inclusion of fixed effects

In this section, I assess simulation results with the inclusion of classroom fixed effects. Griliches (1977) shows that the inclusion of fixed effects will exacerbate attenuation bias in the mismeasured regressor of interest. However, less is known regarding how fixed effects affect estimation bias in other regressors.

Table B2 provides simulation results looking at how class fixed effects affect $\hat{\beta}$. Across rows, I alter the number of classes and the degree of correlation in student achievement within classes. For the simulation with random assignment, I give individuals an assignment

value $x = \mathcal{N}(0, 1)$. For non-random class assignment that has lower correlation with ability, I give individuals an assignment value $x = A + \mathcal{N}(0, 0.5)$. For non-random class assignment that has lower correlation with ability, I give individuals value $x = A + \mathcal{N}(0, 0.1)$. If there are three classes, I assign individuals to classes C such that:

$$C = \begin{cases} 1 & x < -.43 \\ 2 & -.43 \leq x < .43 \\ 3 & x \geq .43 \end{cases}$$

If there are five classes, I assign individuals to classes C such that:

$$C = \begin{cases} 1 & x < -.84 \\ 2 & -.84 \leq x < .25 \\ 3 & -.25 \leq x < .25 \\ 4 & -.25 \leq x < .84 \\ 5 & x \geq .84 \end{cases}$$

Table B2: Measurement error simulation results using math distributions with the inclusion of class fixed effects

Measurement error source	# of Classes	Class Assignment	β	$\hat{\beta}$	SE
$\tilde{T} = A + \mathcal{N}(0, 0.455)$	0		-0.0004	-0.695	0.003
$\tilde{T} = A + \mathcal{N}(0, 0.455)$	3	Random	-0.0004	-0.695	0.003
$\tilde{T} = A + \mathcal{N}(0, 0.455)$	5	Random	-0.0004	-0.695	0.003
$\tilde{T} = A + \mathcal{N}(0, 0.455)$	3	Lower correlation	-0.0004	-0.480	0.002
$\tilde{T} = A + \mathcal{N}(0, 0.455)$	5	Lower correlation	-0.0004	-0.420	0.002
$\tilde{T} = A + \mathcal{N}(0, 0.455)$	3	Higher correlation	-0.0004	-0.024	0.002
$\tilde{T} = A + \mathcal{N}(0, 0.455)$	5	Higher correlation	-0.0004	-0.012	0.002

Results run using 1,000 simulations, and column reporting $\hat{\beta}$ represents the average estimate over these simulations. The sample size is $n = 100,000$ with $R = 0$ for 64% of observations and $R = 1$ for 34% of observations to match the ratio of White and Black students in my sample.