ASSESSED BY A TEACHER LIKE ME: RACE AND TEACHER ASSESSMENTS

Amine Ouazad
INSEAD
Department of Economics
77300 Fontainebleau
France
amine.ouazad@insead.edu

Abstract

Do teachers assess same-race students more favorably? This paper uses nationally representative data on teacher assessments of student ability that can be compared with test scores to determine whether teachers give better assessments to same-race students. The data set follows students from kindergarten to grade 5, a period during which racial gaps in test scores increase rapidly. Teacher assessments comprise up to twenty items measuring specific skills. Using a unique within-student and within-teacher identification and while controlling for subject-specific test scores, I find that teachers do assess same-race students more favorably. Effects appear in kindergarten and persist thereafter. Robustness checks suggest that: student behavior does not explain this effect; same-race effects are evident in teacher assessments of most of the skills; grading "on the curve" should be associated with lower assessments; and measurement error in assessments or test scores does not significantly affect the estimates.

doi:10.1162/EDFP_a_00136

1. INTRODUCTION

A growing body of research in education and psychology argues that minority students receive less favorable feedback and less praise than do their white peers (Meier, Stewart, and England 1989; Marcus, Gross, and Seefeldt 1991; Casteel 1998; Van Ewijk 2011). The research is usually conducted on small samples, which may cast doubt on the wider applicability of results obtained for particular schools or school districts (i.e., on whether results are externally valid; Carpenter, Harrison, and List 2005). In this paper I use a longitudinal and nationally representative data set to measure whether or not teachers assess same-race students more favorably. Field experiments with nationally representative European data sets have recently measured whether teachers assess minority students more favorably (Hinnerich, Höglin, and Johannesson 2011). In the United States, however, there are no nationally representative data on teachers' perceptions of same-race students' skills. Analysis of the National Educational Longitudinal Study of 1988 suggests that teachers have more favorable perceptions of same-race students (Dee 2005), but in that study the variables used to capture those perceptions (e.g., "constantly inattentive," "frequently disruptive," "rarely completes homework") are measures more of student behavior than of student performance. Hence these data cannot be used to infer a same-race effect because such teacher perceptions are not comparable to test scores.

There is another reason why it is so difficult to measure whether teachers assess same-race students more highly. Even if the researcher has comparable teacher assessments of students and test scores, a finding that teachers give better assessments to same-race students (conditional on test scores) could not be given a causal interpretation owing to possible confounding factors. Causal effects can be estimated if the researcher randomizes the assignment of teachers to students, but such randomization is a long and costly process that is usually performed only for small, nonrepresentative samples.

These considerations leave the researcher in a quandary. On the one hand, randomized samples with comparable teacher assessments and test scores provide convincing evidence that teachers have more favorable perceptions of same-race students' skills, but randomized estimates are typically available only for nonrepresentative samples of students. On the other hand, nationally representative samples usually lack two important features: teacher assessments of student performance that are comparable to test scores, and randomized assignment of teachers to students.[1]

1. Lavy (2004) uses a nationally representative sample to estimate the impact of student gender on grades at the high-school matriculation exam in Israel, but teacher assignments are not randomized. Adding unique teacher identifiers to Lavy (2004) would also allow an identification strategy based on comparisons of teacher assessments and test scores while controlling for teacher effects.

This paper uses a longitudinal, nationally representative data set, the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999, which includes detailed teacher assessments and test scores—in both mathematics and English—in each wave of data collection from kindergarten to grade 5. The teacher assessments are available for both subjects, and there are as many as ten questions on specific skills within each subject in each follow-up (Tourangeau et al. 2009). Given these data, continuous teacher assessments can be compared with test scores.[2] Teachers are not randomly assigned to students. Because the data set follows students through five follow-ups (from kindergarten to grade 5) and includes teacher and student identifiers, however, I am able to estimate the same-race effect on teacher assessments by using a unique within-student (i.e., across grades[3]) and within-teacher identification strategy that controls for student- and teacher-specific confounding factors. The paper also describes several robustness checks, which indicate that: (1) behavior does not explain the reported estimate of the same-race effect on teacher assessments; (2) the same-race effect appears in kindergarten for most skills that are assessed by the teacher; (3) grading on the curve within a classroom would result in lower teacher assessments for same-race students; and (4) measurement error in teacher assessments or in test scores has no significant effect on the point estimates.

The within-student identification strategy yields the following result: a student who moves from a same-race teacher in one grade to a different-race teacher in the next grade encounters a significant drop in teacher assessments.[4] Our second, within-teacher identification strategy compares the teacher assessment of same-race students to the average teacher assessment in the student's classroom. I combine the within-student and within-teacher identification strategies and condition the results on student test scores: being assessed by a same-race teacher increases teacher assessments of student performance by 4 percent of a standard deviation in English and by 7 percent of a standard deviation in mathematics.

I design robustness checks to assess whether these results are consistent with a teacher bias in favor of same-race students. One might object that higher teacher assessments for same-race students reflect behavioral differences. After all, teacher assessments of student performance do reflect, in part,

2. Tourangeau et al. (2009) mention that teacher assessments and test scores measure students' skills within the same broad curricular domains. Section 4 examines teachers' perceptions of students skill by skill—and as early as in kindergarten—for skills that are the most likely to be assessed by test scores; the results are similar (if not stronger) same-race effects.
3. Also, the survey is designed in a way that facilitates test score comparisons across grades. The tests consist of two stages: an initial routing test for student ability, and second-stage tests that include questions common to multiple grades (Tourangeau et al. 2009).
4. All of these results are conditional on student test scores.

student behavior (Sherman and Cormier 1974). The within-student identification strategy used here neutralizes the effect of permanent student behavioral differences, but it cannot control for changes in student behavior that could affect teacher assessments. Because the allocation of teachers to students is not random,[5] behavioral changes that raise teacher assessments may correlate with being assigned to a same-race teacher in the subsequent grade. The data set includes four reliable measures of student behavior that are based on the Social Skills Rating System (Gresham and Elliott 1990). These measures vary both across students and across grades. I do not find that behavioral differences between same-race and other-race students explain the within-student and within-teacher estimates of same-race effects on teacher assessments. Neither do I find that changes in behavior from one grade to the next are associated with the student moving from a same-race (other-race) teacher to an other-race (same-race) teacher.

A second possible objection is that, as measures of student performance, test scores are noisy and therefore may not fully condition for student performance when assessing same-race effects on teacher assessments. In that case, teacher assessments could be higher for same-race students simply because same-race students perform better. Test scores and teacher assessments are highly reliable, but the question is whether a small amount of measurement error would be sufficient to confound the estimate of a same-race effect. This paper calculates the impact of a given amount of measurement error in test scores on the derived estimate of the same-race effect. A test score measurement error of 50 percent would be required to account for the estimated same-race effect.

The third major objection to this paper's findings is that teacher assessments may be an implicit ranking of students within a given classroom rather than measures (e.g., test scores) based on a common scale. I have used a simple statistical framework to show that, because minority students have (on average) lower test scores than white students and because minority and white students tend to be in different classrooms, grading on a curve would lead to higher teacher assessments for minority students—even though minority students have significantly (up to 40 percent of a standard deviation) lower teacher assessments. Grading on a curve also would affect estimates of the same-race effect if peer group composition were correlated with assignment to a same-race teacher. Controlling for peers' average test score in the main specification does not affect my estimate of the same-race effect on teacher

5. For some evidence of nonrandom allocation of teachers to students, see Clotfelter, Ladd, and Vigdor (2005).

assessment. Moreover, assignment to a same-race teacher is not significantly correlated with peers' average test score.

My main finding—that students are assessed more highly by teachers of their own race—is robust to the three objections just detailed. That finding is of particular relevance if teacher assessments are shown to have an effect on student achievement. Identifying the impact of teacher perceptions of student skills on later test scores is difficult, and it has led to a large and somewhat controversial literature in psychology and education (Rosenthal and Jacobson 1968; Jussim 1989; Jussim and Harber 2005). In the so-called Pygmalion experiments, a random subset of students in a small sample of participating schools is typically labeled "bloomers," and the research focus is on estimating the effect of such information on student performance. In this paper's nationally representative data set, I find that previous assessments have a significant impact on later test scores (after conditioning for student effects, teacher effects, and grade effects).[6] In fact, previous teacher assessments are more strongly correlated with later test scores than are previous test scores.

The paper contributes to two separate literatures. First, it belongs to the growing literature that documents same-race effects in a number of other contexts. Price and Wolfers (2010) provide statistical evidence that National Basketball Association referees favor players of their own race. In firms, Giuliano, Levine, and Leonard (2009) found that white, Hispanic, and Asian managers hire more whites and fewer blacks than do black managers. In the data set of Giuliano, Levine, and Leonard (2011), employees have better outcomes when they are the same race as their manager. The main contribution of this paper to that literature is providing evidence of same-race effects on perceptions in education while using a nationally representative data set and novel robustness checks.

In studying teacher perceptions of student skills from kindergarten to grade 5, this paper adds also to the literature on teachers' perceptions of minority students during their early years of schooling. The previous literature on race and student assessment has used data for no earlier than grade 8 (Dee 2004). Racial test score gaps expand rapidly much sooner, however; Fryer and Levitt (2004) document that, between the start of kindergarten and the end of first grade, black students' scores fall by 20 percent of a standard deviation relative to white students with otherwise similar characteristics.

The conclusions reported in this paper should be of particular interest to policy makers. First, teachers as a group are less diverse than the U.S. student population. There is, in particular, a persistent gap between the percentage of

6. I also instrument the previous test score by lagged test scores to avoid biases stemming from regression to the mean (see, e.g., Arellano and Bond 1991).

minority teachers and the percentage of minority students. Numerous papers and reports have suggested improvements in the recruitment and retention of minority teachers (Kirby, Berends, and Naftel 1999; Achinstein et al. 2010; Ingersoll and May 2011). Second, the paper's results suggest that involving teachers in student assessments[7] may affect those assessments in ways that reflect racial perceptions. To ensure fairness, therefore, an assessment system that involves teachers should exhibit an appropriate racial balance among graders. Note also that an interesting area of research suggests that racial perceptions are not fixed and can be significantly altered.[8]

The paper is structured as follows. Section 2 presents the data set and descriptive evidence for higher teacher assessments of same-race students (conditional on test scores). Section 3 presents the within-student and within-teacher identification strategies separately before combining them to obtain the paper's baseline estimate. Section 4 discusses the three major objections as well as two policy implications of our results on teacher assessments. Section 5 concludes.

2. DATA SET AND DESCRIPTIVE EVIDENCE

Structure of the Data Set

The data set is the Early Childhood Longitudinal Study, Kindergarten cohort of 1998 (ECLS-K) from the National Center for Education Statistics, U.S. Department of Education. The data follow a nationally representative sample of 20,000 kindergarten students in fall and spring kindergarten 1998, spring grade 1, spring grade 3, and spring grade 5. About a thousand schools participated.

Overall, the design of the experiment is such that observations are mostly missing at random. Follow-ups have combined procedures to reduce costs and maintain the sample's representativeness. Students who move to another school are randomly subsampled to reduce costs, and new schools and children have been added to the data set to strengthen the survey's representativeness. In the spring of 1999, some of the schools that had previously declined participation were included. The new participating children rendered the cross-sectional sample representative of first-grade children, all of whom were followed in the spring of grades 3 and 5. This paper uses weights

7. For instance, Darling-Hammond and Pecheone (2010) argue that teachers should be integrally involved in the scoring of assessments.
8. Stangor, Sechrist, and Jost (2001) show how informing participants that others hold different beliefs about African Americans changes their beliefs about that group. Lyons and Kashima (2003) suggest that interpersonal communication figures strongly in maintaining stereotypes. An interesting avenue for future research involves examining how colleagues' perceptions may affect a teacher's perceptions—using data as in Jackson and Bruegmann (2009) but instead with teachers' perceptions of student performance.

Table 1. Descriptive Statistics

	Mean	SD	Observations
Observations per Student	6.991	(2.020)	115,950
Observations per Teacher	8.198	(5.914)	115,950
Test Score			
English	50.00	(10.00)	67,885
Mathematics	50.00	(10.00)	48,065
Teacher Assessment			
English	50.00	(10.00)	67,885
Mathematics	50.00	(10.00)	48,065
Teacher Race[a]			
White, non-Hispanic	0.809	(0.393)	115,950
Black, non-Hispanic	0.063	(0.244)	115,950
Asian, non-Hispanic	0.019	(0.135)	115,950
Hispanic, any race	0.052	(0.221)	115,950
Other race, non-Hispanic	0.057	(0.232)	115,950
Student Race[a]			
White, non-Hispanic	0.587	(0.492)	115,950
Black, non-Hispanic	0.137	(0.344)	115,950
Asian, non-Hispanic	0.057	(0.232)	115,950
Hispanic, any race	0.157	(0.364)	115,950
Other race, non-Hispanic	0.062	(0.241)	115,950
Same-race Teacher by Student Race[b]	0.436	(0.496)	115,950
White, non-Hispanic	0.683	(0.465)	115,950
Black, non-Hispanic	0.188	(0.391)	115,950
Asian, non-Hispanic	0.069	(0.253)	115,950
Hispanic, any race	0.163	(0.369)	115,950
Other race, non-Hispanic	0.056	(0.230)	115,950

[a]Other race, non-Hispanic includes Pacific Islanders, American Indians, and non-Hispanic students reporting multiple races.
[b]Both of the same race, non-Hispanic, or Hispanic, any race.

provided by the survey's designers to estimate representative effects, though the analysis is robust to changes in weights.

Observations that lacked data on basic variables (test scores, subjective assessments, teachers' and children's race and gender) were deleted.[9] The analysis in this paper is based on 48,065 observations in mathematics and 67,885 in English, numbers that are similar to Fryer and Levitt (2006).

The restricted-use version of the data set includes both student and teacher identifiers. Hence, students can be followed across grades. Within each follow-up, observations can be grouped by classroom using the teacher identifiers. Table 1 shows that data set includes about 6.9 observations per student (3.45

9. Results are robust to an alternative specification where missing observations are present with a dummy variable indicating that the data are missing.

on average per student in each subject); the data set includes 8.2 observations per teacher.

Test Scores and Teacher Assessments

Test scores are based on answers to multiple-choice questionnaires conducted by external assessors. They conform to national and state standards.[10] Overall, tests ask more than seventy questions in English, and more than sixty questions in mathematics. Skills covered by the English assessments from kindergarten to fifth grade include: print familiarity, letter recognition, and beginning and ending sounds; recognition of common words (sight vocabulary) and decoding multisyllabic words; vocabulary knowledge, such as receptive vocabulary and vocabulary in context; and passage comprehension. Skills covered by the mathematics assessment include: number sense, properties, and operations; measurement; geometry and spatial sense; data analysis, statistics, and probability; and patterns, algebra, and functions. Test scores were standardized to a mean of 50 and a standard deviation of 10 (table 1). Reliability measures based on repeated estimates of test scores indicate that the tests are highly reliable; Rasch coefficients range between 0.88 and 0.95, inclusive.

Teacher assessments of student skills[11] are collected at approximately the same time as the tests are taken. Up to the spring of grade 3, the same teacher in English and in mathematics assesses students. A different teacher assesses students in each grade. Teachers do not see the test results, so that test score results do not directly affect teacher assessments. The user guide specifies that "This is not a test and should not be administered directly to the child" (see, for example, the Spring 2004 Fifth Grade questionnaire[12]). Teachers complete one questionnaire per student. There are three different teacher assessments: for language and literacy, mathematical thinking, and general knowledge. The current paper uses the English (language and literacy) and mathematics (mathematical thinking) assessments, as there is no corresponding test score for general knowledge. The instructions make it clear that these assessments should not be administered as a test directly to the student. For English and for mathematics, teachers answer seven to nine questions, for a total number of fourteen to eighteen questions. Answers are on a 5-point scale: Not Yet,

10. These include the National Assessment for Educational Progress, the National Council of Teachers of Mathematics, the American Association for the Advancement of Science, and the National Academy of Sciences.
11. In the ECLS-K user guide, teacher assessments are also known as the academic rating scale.
12. Page 3 of the 2004 Grade 5 mathematics form: "Please rate this child's skills, knowledge, and behaviors in mathematics based on your experience with the child identified on the cover of this questionnaire. This is NOT a test and should not be administered directly to the child. Each question includes examples that are meant to help you think of the range of situations in which the child may demonstrate similar skills and behaviors."

Beginning, In Progress, Intermediate, and Proficient. An overall assessment is computed for English and for mathematics. Teacher assessments, like test scores, were standardized to a mean of 50 and a standard deviation of 10 (table 1). Reliability measures suggest that teacher assessments are highly reliable; Rasch coefficients range between 0.87 and 0.94.

Descriptive Evidence of Same-Race Effects on Teacher Assessments

The restricted-use version of the ECLS-K reports teachers' and students' race and gender. The survey combines race and ethnicity for teachers. "Hispanic, any race" is one category, and others are "White, any race," "Black, any race," and so on. The survey does distinguish race and ethnicity for students, however. The two variables for students' race and ethnicity were hence combined to match the single teacher's race and ethnicity variable. Hence "same race" should be read as "same race (non-Hispanic) or both Hispanic (any race)."[13]

The data set oversamples students from racial and ethnic minorities to increase the precision of the estimates. In the data set, 14 percent are black students, 16 percent are Hispanic students, and 6 percent are Asian students. There are significantly more white teachers than white students as a fraction of the observations, and significantly fewer black, Hispanic, and Asian teachers compared with the corresponding fractions of black, Hispanic, and Asian students. Hence a white student is significantly more likely to be assessed by a same-race teacher than a black, Hispanic, or Asian student.

Figure 1 presents the average teacher assessments at each test score level, for students assessed by a same-race teacher and for students assessed by a teacher of another race. Each line is a local polynomial regression of teacher assessments on test scores;[14] the solid line (the dashed line) is estimated on observations for students assessed by a same-race teacher (a teacher of another race). The two graphs suggest that, at most test score levels, students have on average higher teacher assessments when assessed by a same-race teacher. The gap appears larger for Hispanic students (bottom graph) than for black students (top graph).

13. Also the student's race variable follows the 1997 U.S. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity published by the Office for Management and Budget, which allow for the possibility of specifying "more than one race." However, the share of multiracial students is small. Multiracial students are classified as "Other race," but results are robust to alternative classifications.
14. Figure generated with local mean smoothing with 500 points, Epanechnikov kernel, and optimal half-width. The gap is robust to a variety of number of points, kernels, and half-width sizes.

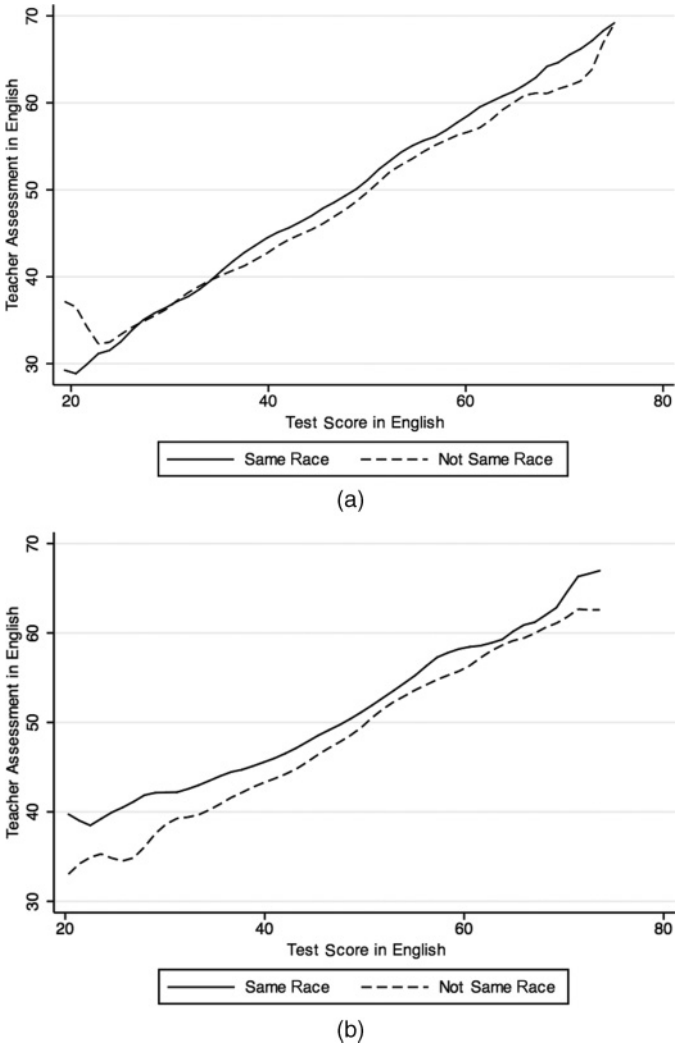

Figure 1. Descriptive Evidence of the Same-Race Effect in (a) Black Students and (b) Hispanic Students. *Notes:* Each panel plots a local polynomial regression of teacher assessments on test scores, using an Epanechnikov kernel, 500 points, and optimal half-width. The gap between the two curves is present even when changing the type of the kernel, number of points, and the half-width.

An ordinary least squares (OLS) regression estimates the average effect of same race teachers on the difference between teacher assessments and test scores, and provides confidence intervals:

$$\begin{aligned} TA_{i,f,t} - TS_{i,f,t} = \text{constant} &+ \delta \cdot Same\,Race_{i,f,t} \\ &+ StudentCharacteristics_i \beta \\ &+ TeacherCharacteristics_{i,f,t} \gamma \\ &+ Grade_t + \varepsilon_{i,f,t} \end{aligned} \tag{1}$$

where i indexes students, f the subject area (mathematics or English), and t the wave of the longitudinal data ($t = \{$Fall kindergarten, spring kindergarten, spring grade 1, spring grade 3, spring grade 5$\}$). $TA_{i,f,t}$ is the standardized teacher assessment, $TS_{i,f,t}$ represents the standardized test score. $Same\ Race_{i,f,t}$ is a dummy set to 1 if student i in subject f in wave t was assessed by a same-race teacher. $Student\ characteristics_i$ is a vector of dummies for the student's gender and race. $Teacher\ Characteristics_{i,f,t}$ is a vector of dummies for student i's teacher in subject f in wave t. $Grade_t$ is a grade effect, and $\varepsilon_{i,f,t}$ is the residual, clustered by student.[15]

The regression is performed separately for English and for mathematics. Throughout the paper, I also present the regression with the teacher assessment as the dependent variable, and the test score as a control. While the regression with the test score as an explanatory variable corresponds to the concept of conditional bias (Ferguson 2003), putting the test score on the right-hand side means that the estimate of the coefficient of the same-race dummy may capture measurement error in test scores. Specification 1 has both teacher assessment and test score on the left-hand side, which substantially alleviates any bias caused by measurement error.

The OLS regression suggests that a student assessed by a same-race teacher gets a teacher assessment that is about 2.8 percent to 5.7 percent of a standard deviation higher in mathematics, and 4.3 percent to 6.7 percent of a standard deviation higher in English (table 2). In this specification, the test score as an explanatory variable explains only 34.8 to 44 percent of the variance of teacher assessments.

3. IDENTIFICATION STRATEGY

Within-Student Identification: Using Student Mobility from/to a Same-Race Teacher

In the descriptive evidence that was presented in the previous section, the OLS estimate of the same-race effect may be biased because a number of student-specific variables are omitted from the regression.

For instance, literature suggests that teacher perceptions of student performance might depend on a number of characteristics other than student race: student behavior (Sherman and Cormier 1974), language (Gluszek and Dovidio 2010), parental involvement (Wilson and Martinussen 1999), student academic engagement (Hughes and Kwok 2007), and other factors. Neither of these variables is measured by test scores nor reflects racial perceptions per se.

15. Clustering by classroom, by student, or two-clustering (Cameron, Gelbach, and Miller, 2011) by both student and classroom has little impact on the standard errors. Because two-way clustering with two-way fixed effect (used later in section 3) does not yet exist in the literature, I chose to present standard errors clustered by student. Clustering by classroom yields very similar standard errors in all specification.

Table 2. OLS Regressions

	Mathematics		English	
	(1)	(2)	(3)	(4)
	Teacher Assessment	Teacher Assessment – Test Score	Teacher Assessment	Teacher Assessment – Test Score
Same-Race	0.281*	0.566**	0.428**	0.665**
	(0.118)	(0.131)	(0.093)	(0.122)
Test Score	0.591**	–	0.659**	–
	(0.004)		(0.003)	
Controls	Student and teacher race and gender, grade effects			
Observations	48,065	48,065	67,855	67,855
Students	20,252	20,252	20,252	20,252
Teachers	5,297	5,297	5,496	5,496
R^2	0.348	0.034	0.436	0.029
F Statistic	1,218.5	85.3	2,501.1	68.9

Notes: Standard errors clustered by student. Clustering by classroom yields similar significance levels. Test scores and teacher assessments are standardized to a mean of 50 and a standard deviation of 10.
*Statistically significant at the 5% level; **statistically significant at the 1% level.

Identifying the specific effect of the student's race requires a more complete specification than equation 1, one that at least controls for student-specific omitted variables. Such omitted variables will confound the estimate of the same-race effect if teachers and students are non-randomly matched.

Assume that the teacher assessment incorporates a measure of the test score, captures a same-race bias, and also student-specific omitted variables:

$$TA_{i,f,t} = \text{constant} + \delta \cdot Same\,Race_{i,f,t} + \alpha Test\,Score_{i,f,t} + Grade_t + Controls_{i,f,t} + Student\,Omitted\,Variable_{i,f,t} + Residual_{i,f,t} \quad (2)$$

with the same notations as in specification 1, and $\varepsilon_{i,f,t} = Student\ Omitted\ Variable_{i,f,t} + Residual_{i,f,t}$. $Controls_{i,f,t}$ is a set of dummies for the teacher's race and gender. If student-specific omitted variables that have a positive impact on the teacher assessment are correlated with assignment to a same-race teacher, the effect δ of a same-race teacher on assessments is overestimated. In other words, if assignment to teachers depends on unobservables that affect teacher assessments, the same-race effect is biased. Student-specific omitted variables that are not correlated with same-race assignments will also imply a correlation of residuals common to a given student, that is, $\text{Corr}(\varepsilon_{i,f,t},\varepsilon_{i,f',t'})$ is

not equal to 0, and standard errors will need to be corrected for student-level clustering.[16]

If student-specific omitted variables do not vary across grades,[17] specification 2 can be estimated using a student fixed effect $Student_{i,f}$:

$$TA_{i,f,t} = \text{constant} + \delta \cdot Same\,Race_{i,f,t} + \alpha \cdot Test\,Score_{i,f,t} + Controls_{i,f,t} + Student_{i,f} + Grade_t + Residual_{i,f,t} \quad (3)$$

which is estimated using either a set of student dummies, or in first-difference. A major advantage of the dummy variable approach is that it allows us to recover an estimate of the student unobservables $Student_i$; using this estimate we can check whether there is a significant correlation between assignment to a same-race teacher and student unobservables. Specification 3 can also be estimated in first-difference,[18] that is, using a within-student regression:

$$\begin{aligned} TA_{i,f,t+1} - TA_{i,f,t} = {} & \delta(Same\,Race_{i,f,t+1} - Same\,Race_{i,f,t}) \\ & + (Controls_{i,f,t+1} - Controls_{i,f,t}) \\ & + \alpha(Test\,Score_{i,f,t+1} - Test\,Score_{i,f,t}) \\ & + (Grade_{t+1} - Grade_t) + (Residual_{i,f,t+1} \\ & - Residual_{i,f,t}). \end{aligned} \quad (4)$$

The first-differenced specification makes clear that the identification of the same-race effect δ relies on student mobility from/to a same-race teacher. The effect of a same-race teacher is estimated without bias if the mobility of a student from a teacher of the same-race (another-race) in one grade to a teacher of another race (the same race), in the next grade, is uncorrelated with time varying student unobservables that have an impact on test scores, that is, $\text{Corr}((Same\,Race_{i,f,t+1} - Same\,Race_{i,f,t}), (Residual_{i,f,t+1} - Residual_{i,f,t})) = 0$. Student behavior is one such time varying unobservable that may affect teacher assessments and is potentially correlated with student mobility to/from

16. Specifically, $\text{Cov}(\varepsilon_{i,f,t},\varepsilon_{i,f',t'}) = \text{Cov}(Student\,Omitted\,Variable_{i,f,t}, Student\,Omitted\,Variable_{i,f',t'})$ for $f \neq f'$ and for $t \neq t'$. If student-specific omitted variables are constant across grades, then $\text{Cov}(\varepsilon_{i,f,t}, \varepsilon_{i,f,t'}) = \text{Var}(Student\,Omitted\,Variable_{i,f})$ and the correlation of residuals for a given student across grades will be equal to the ratio of the variance of student unobservables to the overall variance of the residuals (Moulton 1990).
17. $Student\,Omitted\,Variable_{i,f,t} = Student\,Omitted\,Variable_{i,f,t'}$ for any t, t'.
18. Both approaches (student dummies and first-differenced specification) are equivalent with a large number of observations as long as the strict exogeneity assumption is satisfied (Baltagi 2008), that is, $E(Residual_{i,f,t}|X_{i,f,1}, X_{i,f,2}, \ldots, X_{i,f,5}) = 0$, where 1,2,...,5 indexes waves of the survey, and $X_{i,f,t}$ denotes the vector of explanatory variables for student i in subject area f, in grade t (the constant, same race dummy, test score, and grade dummies).

a teacher of the same race. I discuss the impact of behavior on estimates in section 4.

Because identification relies on student mobility across teachers, it is important to check that a sufficient number of students move to teachers of different races. Otherwise identification would rely on a small number of students who move from/to a teacher of the same race.[19] There are a large number of such moves: 51 percent of students experience mobility from/to a same-race teacher at some point between kindergarten and grade 5, and the sample of movers is balanced in terms of race, gender, and parental income.[20]

Columns (1) and (4) of table 3 present the estimation of the first-differenced specification 4 in mathematics and in English, with standard errors clustered by student.[21] Being assessed by a teacher of the same race raises teacher assessments by 3.5 percent of a standard deviation in mathematics and by 4.3 percent in English. The specification has fewer observations because the number of observations is equal to the number of first-differenced teacher assessments. Columns (2) and (5) present results of the estimation of specification 3, which includes a student fixed effect. Being assessed by a teacher of the same race raises assessments by 7 percent of a standard deviation in mathematics and by 4.8 percent of a standard deviation in English. The regression is strongly significant with an F statistic of 82.6. Importantly, there is a significantly positive correlation between the estimated student effects and assignment to a same race teacher both in mathematics and in English, which indicates that the regression without student fixed effects underestimates the impact of a same-race teacher on assessments. Columns (3) and (6) regress the difference between the teacher assessment and the test score on the explanatory variables. Estimates of the same race effect are comparable to columns (2) and (5) of the same table.

Within-Classroom Identification

Teacher-specific omitted variables may also confound the estimate of the same-race effect. Although OLS specification 1 controls for teachers' race and gender, other teacher characteristics, imperfectly correlated with race and gender, affect teacher assessments. For instance, Figlio and Lucas (2004) find that some teachers give higher average grades regardless of their students' ability, race, or gender. Such variation in average assessments across classrooms should

19. In general, if a covariate does not vary for a given student in a panel data regression with student fixed effects, the student's observation will not contribute to the estimation of the effect (Wooldridge 2002).
20. At each parental income level, from 41 percent to 52 percent of students experience a transition from/to a same race teacher. Statistics available on request.
21. Clustering either by classroom, by student, or clustering by both classroom and student (Cameron, Gelbach, and Miller 2011) does not significantly affect the estimated standard errors.

Table 3. Results with First-Differenced Specification and with Student Fixed Effects

	Mathematics			English		
	(1)	(2)	(3)	(4)	(5)	(6)
	First-Differenced Teacher Assessment	Teacher Assessment	Teacher Assessment – Test Score	First-Differenced Teacher Assessment	Teacher Assessment	Teacher Assessment – Test Score
Same-Race	0.350^{+}	0.704**	0.784**	0.429**	0.413**	0.483**
	(0.211)	(0.162)	(0.179)	(0.154)	(0.113)	(0.176)
Test Score	0.129**	0.263**	–	0.241**	0.316**	–
	(0.011)	(0.009)		(0.007)	(0.006)	
Student Effect	No	Yes	Yes	No	Yes	Yes
Student and Teacher Race and Gender	Yes	No	No	Yes	No	No
Grade Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	22,089[a]	48,065	48,065	44,492[a]	67,855	67,855
R^2	0.010	0.665	0.040	0.036	0.699	0.430
F Statistic for Student Effects (*p* value)	–	3.108	2.372	–	2.024	1.646
		(0.000)	(0.000)		(0.000)	(0.000)
Corr(*Same Race, Student Effects*)	–	0.042**	−0.145**	–	0.065**	−0.096**
		(0.000)	(0.000)		(0.000)	(0.000)

Notes: Standard errors clustered by student. Results robust to clustering by classroom. Test scores and teacher assessments standardized to a mean of 50 and a standard deviation of 10.

[a]Smaller number of observations due to first differencing.

*Statistically significant at the 5% level; **statistically significant at the 1% level; $^{+}$statistically significant at the 10% level.

be controlled for in specification 1 as the nonrandom sorting of teachers to students implies that the teacher's average assessment may be correlated with assignment to a same-race student.

All these teacher-specific omitted variables enter in the determination of teacher assessments:

$$\begin{aligned} TA_{i,f,t} = \text{constant} &+ \delta \cdot Same\,Race_{i,f,t} + \alpha\, Test\,Score_{i,f,t} \\ &+ Teacher\,Omitted\,Variable_{i,f,t} + Controls_{i,f,t} \\ &+ Grade_t + Residual_{i,f,t}. \end{aligned} \tag{5}$$

Teacher omitted variables ($Teacher\ Omitted\ Variable_{i,f,t}$), if correlated positively with assignment to a same race teacher ($Same\ Race_{i,f,t}$), lead to an upward bias in the estimate δ of the same-race effect. The presence of teacher-specific omitted variables also imply a correlation of residuals in the OLS specification across observations of the same classroom, and standard errors should be corrected for clustering at the classroom level.[22] Because of the large number of fixed effects (6,093 teachers), a specification like specification 5 is usually estimated by taking the within-classroom difference of teacher assessments, test scores, and each covariate of the specification

$$\begin{aligned} & TA_{i,f,t} - E(TA_{.,f,t}|classroom) \\ = \delta \cdot & (Same\,Race_{i,f,t} - E(Same\,Race_{i,f,t}|classroom) \\ & + \alpha \cdot (TS_{i,f,t} - E(TS_{.,f,t}|classroom)) \\ & + Controls_{i,f,t} - E(Controls_{i,f,t}|classroom)) \\ & + Residual'_{i,f,t}. \end{aligned} \tag{6}$$

where $E(x_{.,f,t}|classroom)$ is the average of covariate x in the classroom of student i in subject f in year t. The within-classroom specification makes it clear that the identification relies on comparing the teacher assessment $TA_{i,f,t}$ of a student to the average teacher assessment $E(TA_{.,f,t}|classroom)$ in the classroom. A classroom contributes to the identification of the same-race effect if it has both same-race and other-race students.[23] Fortunately, 97.2 percent of the classrooms of the sample have observations of same-race and other-race students, and 44 percent of students are of the same race as teacher on average.

22. Throughout the paper I cluster standard errors at the student level, but clustering at the classroom level or two-way clustering at the student and classroom levels (Cameron, Gelbach, and Miller 2011) yields similar significance levels.

23. Formally, if the value of $Same\ Race_{i,f,t} - E(Same\ Race_{.,f,t}|classroom)$ changes within a classroom.

Specification 5 can also be estimated by including a set of teacher fixed effects, namely, one dummy of each teacher of the sample.

$$TA_{i,f,t} = \text{constant} + \delta \cdot Same\,Race_{i,f,t} + \alpha Test\,Score_{i,f,t} + Teacher\,Effect_{i,f,t} + Grade_t + Residual_{i,f,t}. \quad (7)$$

Both approaches (specifications 6 and 7) yield the same estimate with a large number of observations (Baltagi 2008).[24] The advantage of such a specification is that it allows us to recover an estimate of the teacher effect. In all waves except the spring grade 5 follow-up, the same teacher assesses students in English and mathematics, but separate teacher effects are estimated for English and for mathematics.

Columns (1) and (4) of table 4 show the results of the within-classroom specification 6. Students assessed by a teacher of the same race have higher teacher assessments, by 4.1 percent of a standard deviation in English and 5.5 percent in mathematics. All results are significant at 1 percent. Interestingly, test scores and observable controls explain 34 percent of the variance of teacher assessments. Columns (2) and (5) present results of the estimation of specification 7, which includes teacher effects. The point estimates are larger than in the within-teacher approach, but they are not statistically different from the estimates of columns (1) and (4). Having a same-race teacher raises teacher assessments by 6.9 percent of a standard deviation in English and 7.0 percent of a standard deviation in mathematics. The specification allows us to estimate that teacher effects are significant (the null hypothesis that teacher effects are equal to zero is rejected), indicating that teacher unobservables play a role in assessments. Moreover, being assessed by a same-race teacher is negatively correlated with the teacher effect (especially in mathematics), and we indeed observe a downward bias: The OLS estimation of the same-race effect without teacher effects in columns (1) and (3) of table 2 is lower than the estimates of columns (2) and (5) of table 4. Finally, results available on request show that teacher unobservables are not accounted for by the teacher's race, gender, experience, or tenure.

Combining the Within-Student and Within-Classroom Identification Strategies

Finally, I combine both the former two identification strategies to control for both student-specific and teacher-specific omitted variables. My preferred

24. That is, both estimators converge in probability to the same estimate. Under the assumption that residuals are strictly exogenous within each classroom, that is, $E(Residual'_{i,f,t}|X_{\cdot,f,t}) = 0$, where $X_{i,f,t}$ is the vector of explanatory (right-hand side) variables in specification 6.

Table 4. Results of the Within-Teacher Estimation, of the Teacher Fixed-Effects Specification, and Combining both Student and Teacher Fixed Effects

	Mathematics			English		
	(1)	(2)	(3)	(4)	(5)	(6)
	Teacher Assessment – Average TA	**Teacher Assessment**	**Teacher Assessment**	**Teacher Assessment – Average TA**	**Teacher Assessment**	**Teacher Assessment**
Same Race	0.406**	0.694**	0.711**	0.549**	0.702**	0.435**
	(0.119)	(0.120)	(0.190)	(0.098)	(0.094)	(0.114)
Test Score	0.565**	0.588**	0.241**	0.654**	0.669**	0.313**
	(0.005)	(0.004)	(0.009)	(0.004)	(0.003)	(0.005)
Teacher Effects	No	Yes	Yes	No	Yes	Yes
Student Effects	No	No	Yes	No	No	Yes
Student and Teacher Observables	Yes	Yes	No	Yes	Yes	No
Observations	48,065	48,065	48,065	67,855	67,855	67,855
R^2	0.338	0.540	0.786	0.438	0.553	0.773
Teacher Effects *F* Stat. (*p* value)	–	3.291	2.996	–	2.836	2.786
		(0.000)	(0.000)		(0.000)	(0.000)
Corr(*Same Race, Teacher Effects*)	–	–0.011**	0.020**	–	–0.017**	0.013**
		(0.018)	(0.000)		(0.000)	(0.000)
Student Effects *F* Stat. (*p* value)	–	–	1.794	–	–	2.152
			(0.000)			(0.000)
Corr(*Same Race, Student Effects*)	–	–	0.030**	–	–	0.058**
			(0.000)			(0.000)

Notes: All specifications include grade effects. Standard errors clustered by student. Clustering by classroom yields similar estimates. TA = teacher assessment.
**Statistically significant at the 1% level.

estimate is thus the same-race δ coefficient in the regression that controls for both teacher effects and student effects:

$$\begin{aligned} TA_{i,f,t} = \text{constant} &+ \delta\, Same\, Race_{i,f,t} + \alpha\, Test\, Score_{i,f,t} \\ &+ Teacher\, Effect_{i,f,t} + Student\, Effect_i \\ &+ Grade_t + Residual_{i,f,t} \end{aligned} \tag{8}$$

where the teacher effect (*Teacher Effect*$_{i,f,t}$) and the student effect (*Student Effect*$_i$) are estimated by including a set of dummies for teachers and a set of dummies for students as controls. The large number of students (21,409) and the large number of teachers (6,093) make it necessary to estimate the model using econometric techniques pioneered by Abowd, Creecy, and Kramarz (2002) and Abowd, Kramarz, and Margolis (1999) in the labor economics employer–employee literature. The technique provides estimates for all student effects, teacher effects, grade effects, and same-race and test score coefficients. Standard errors are clustered at the student level; clustering by classroom yields similar standard errors.

Columns (3) and (6) present the estimates. Teachers give better assessments to students of their own race; the effect is 7.1 percent of a standard deviation in mathematics and 4.4 of a standard deviation in English. Teacher and student effects are significant.

4. DISCUSSION OF THE FINDINGS

Behavior and Assessments

Teacher assessments of student performance are partly determined by student behavior (Sherman and Cormier 1974). Column (1) (respectively, Column (2)) of table 5 shows a regression of mathematics teacher assessments (respectively, English teacher assessments) on four behavioral measures.

The four behavioral measures come from a separate questionnaire of each wave of the study. Teachers reported the measures in terms of the social rating scale: approaches to learning, interpersonal skills, externalizing problems behavior, internalizing problems behavior. The scale for approaches to learning measures the ease with which children can benefit from their learning environment. The interpersonal skills scale rates the child's skill in forming and maintaining friendships; getting along with people who are different; comforting or helping other children; expressing feelings, ideas, and opinions in positive ways; and showing sensitivity to the feelings of others. The externalizing problem behaviors scale (i.e., impulsive/overactive scale) addresses acting-out behaviors, and the internalizing problem behavior scale addresses evidence of anxiety, loneliness, low self-esteem, or sadness.

Table 5. Behavior and Assessments

	(1) Mathematics Teacher Assessment	(2) English Teacher Assessment	(3) Same Race	(4) Same Race
Same Race	0.707**	0.419**		
	(0.199)	(0.134)		
Test Score	0.207**	0.265**	–0.001	0.001
	(0.008)	(0.006)	(0.001)	(0.000)
Approaches to Learning	0.267**	0.298**	0.001**	–0.001+
	(0.008)	(0.004)	(0.001)	(0.000)
Interpersonal Skills	0.042**	0.035**	–0.001	0.000
	(0.007)	(0.004)	(0.001)	(0.000)
Externalizing Problem Behavior	0.045**	0.035**	–0.001	0.001
	(0.012)	(0.003)	(0.001)	(0.001)
Internalizing Problem Behavior	–0.040**	–0.058**	–0.001**	0.001**
	(0.006)	(0.005)	(0.000)	(0.000)
Student and Teacher Race and Gender	No	No	Yes	Yes
Student Effects	Yes	Yes	No	Yes
Teacher Effects	Yes	Yes	No	No
F Statistic	4.62		4,249.2	26.66
R^2	0.73	0.80	0.59	0.79
Observations	48,065	67,855	67,855[a]	67,855[a]

[a]Regression performed using English observations. Students are assessed by the same teacher in English and mathematics from kindergarten to grade 3, and different teachers in grade 5. Similar results hold when estimating the regression with mathematics observations.
**Statistically significant at the 1% level. All specifications include grade effects. Standard errors clustered by student. Clustering by classroom yields similar estimates.

The measures of behavior vary substantially, both across students and for a given student, across time. On the interpersonal skills scale, 50.1 percent of the variance is explained by within-student variance, and the behavioral measure in the previous wave of the study explains about 31 percent of the variance of the behavioral measure of the next grade.

In Column (1) of table 5, the teacher assessment in mathematics is regressed on the mathematics test score, the same-race dummy, the four behavioral measures, a student effect, and a teacher effect.

The first noticeable fact is the impact of behavior on assessment. Smaller values indicate stronger behavioral problems. A one standard deviation increase in the approaches to learning scale raises teacher assessments by 3 percent of a standard deviation. A one standard deviation increase in the interpersonal skills measure raises teacher assessments by 0.4 percent of a standard

deviation. Externalizing behavior problems has a similar positive effect. Internalizing behavior problems has a negative impact on teacher assessments. That last result is consistent with the finding (Rutherford, Quinn, and Mathur 2004) that students with internalizing behavior problems (social withdrawal, anxiety, depression) are harder to identify than students with externalizing behavior problems (noncompliance, aggression, disruption).

How behavior affects the baseline estimate of the same-race effect in specification 8 depends on whether students are partly matched to teachers based on their behavior. Because I am using a student fixed-effect regression, behavior is a confounding factor in the regression if changes in behavior across grades are significantly correlated with the probability of being assigned a same-race teacher. If students whose behavior improves are more likely to be assigned to a same-race teacher, the same-race effect δ in specification 8 will be overestimated. Column (3) regresses the same-race dummy on the test score, the four behavioral measures, and student and teacher race and gender dummies. The effect of behavior on same-race assignments is either nonsignificant or very small. Column (4) confirms the finding when including student fixed effects.

Unsurprisingly, therefore, behavioral controls leave the same-race effect (0.707 compared with 0.702 in mathematics, 0.420 compared with 0.435 in English) virtually unchanged compared with the estimate with a student effect and a teacher effect in table 4.

Same-Race Effects Skill by Skill

Table 6 presents results of baseline regression for English, considering only kindergarten fall semester observations. The novelty is that the dependent variable is the teacher assessment broken down into eight separate skills. The results are informative with regard to the likelihood of a bias for two reasons: First, it is unlikely that students benefit from the better teaching of a same-race teacher (Dee 2005) only a few weeks after the start of school and hence better teacher assessments for same-race students are more likely to represent perceptions rather than actual skills. Second, same-race assessment gaps appear also for the least abstract questions—in other words, questions that address the skills that are most likely to be captured by achievement tests.

Take, for example, the statement: "This child easily and quickly names all upper- and lower-case letters of the alphabet." In the fall semester of kindergarten, teachers assess students of their own race 4 percent of a standard deviation higher than children of other races. This English skill is measured in the kindergarten test and is measured early in the curriculum. And similar regressions in grade 5 present similar positive same-race effects.

The same-race effect can also be estimated separately for each grade by including interactions between the grade dummies and the same-race dummy.

Table 6. Same Race Effects Skill by Skill

	Fall Kindergarten English Teacher Assessments							
	Complex	**Understands**	**Names**	**Rhyming**	**Reads**	**Writing**	**Conventions**	**Computer**
Same-Race Teacher	1.257**	1.035**	0.397**	0.674**	0.080	0.018	0.136	0.196*
Same-Race Teacher	(0.142)	(0.146)	(0.127)	(0.106)	(0.104)	(0.108)	(0.098)	(0.077)
Controls	English Test Scores and Teacher Effects							
Observations	16,864	16,864	16,864	16,864	16,864	16,864	16,864	16,864
R^2	0.67	0.65	0.74	0.82	0.83	0.82	0.85	0.91
F Statistic	2,039.7	2,192.3	4,565.3	1,827.2	1,123.4	1,529.9	1,045.9	388.3

Notes: Test scores have a standard deviation of 10 and a mean of 50; child controls include controls for race and gender; teacher controls include controls for the teacher's race, gender, tenure, and experience.
Definitions: Complex = This child uses complex sentence structures. Understands = This child understands and interprets a story or other text read to him/her. Names = This child easily and quickly names all upper- and lower-case letters of the alphabet. Rhyming = This child produces rhyming words. Reads = This child reads simple books independently. Writing = This child demonstrates early writing behaviors. Conventions = This child demonstrates an understanding of some of the conventions of print. Computer = This child uses the computer for a variety of purposes.
*Statistically significant at the 5% level; **statistically significant at the 1% level.

These results (available from the author) show that teachers give more favorable assessments to same-race students as soon as in the fall of kindergarten: 14 percent of a standard deviation higher in mathematics and 11 percent of a standard deviation higher in English. After the fall semester of kindergarten, the effect is about 6 percent (3 percent) of a standard deviation in mathematics (English).

Measurement Error in Test Scores and Teacher Assessments

Two types of measurement error may confound the main estimates of our same-race effect in specification 3. First, teacher assessments may be noisy measures of teacher perceptions of student performance. Second, test scores of multiple-choice questionnaires may be noisy measures of underlying ability (Rudner and Schafer 2001). Random error may be introduced in the design of the questionnaire and distractors (wrong options) may be partially correct. Measurement error in test scores may also be due to the student's sleep patterns, illness, and careless errors when filling out the questionnaire, misinterpretation of test instructions, and other exam conditions.

Measurement error in teacher assessments is likely to make our estimates of the same-race effect less significant, because classical measurement error on the dependent variable of a linear regression (specification 3) does not typically bias estimates but leads to larger standard errors for the estimated coefficients (Wooldridge 2002; Greene 2011). Hence, finding a significant effect of a same-race teacher is evidence that teacher assessments are a sufficiently precise[25] measure of teacher perceptions of student performance.

Measurement error in test scores may be more problematic. Indeed, proper conditioning for student ability in a given grade is key to the estimation of same-race effects on teacher perceptions of students' skills. This paper measures conditional bias as in Ferguson (2003)—that is, the impact of the student's race on teacher assessments when conditioning on covariates that include measures of student ability. The main specification (specification 8) estimates same-race effects on teacher assessments conditional on test scores and student effects. At the extreme, if test scores are such a noisy measure of student ability that most of its variance is accounted for by measurement error, conditioning on test scores will have no impact on the same-race coefficient; the coefficient on test scores will be nonsignificant.[26] In such a case, the same-race coefficient will measure a sum of the same-race effects on teacher perceptions

25. Precision in the statistical sense, as the inverse of the standard deviation.
26. In table 4, the coefficient for test scores in all regressions is less than 1, whereas we would naturally expect this coefficient to equal to 1, given that both assessments and test scores have a standard deviation of 10. Constraining this coefficient to be equal to 1 does not significantly alter the coefficients of interest. Results available on request.

and the positive effect of same-race teachers on student ability (Dee 2005). On the other extreme, if test scores measure student ability accurately,[27] the same-race coefficient in specification 9 will be an estimate of same-race biases.

ECLS-K documentation specifies that test scores are highly reliable (see section 2). But the question here is whether a small amount of measurement error in test scores can explain away the same-race effect—that is, if the same-race coefficient captures some unobserved student ability rather than a bias in teacher assessments.

So is there some amount of measurement error that explains the same-race estimates of table 4? Test scores are noisy measures of the child's underlying ability, so that $Test\ score_{i,f,t} = Ability_{i,f,t} + \nu_{i,f,t}$. Measurement error is assumed to be classical (i.e., $\nu_{i,t}$ is not correlated with ability), which, as Bound, Brown, and Mathiowetz (2001) suggest, is a reasonable assumption in many common cases.

Assume also that teacher assessments capture student ability and are affected by a same- race bias δ:

$$TA_{i,f,t} = \text{constant} + \alpha Student\ Ability_{i,f,t} + \delta Same\ race_{i,f,t} + \varepsilon_{i,f,t}. \quad (9)$$

For clarity and without loss of generality, student and teacher fixed effects are not included in this equation. I do not observe student ability and so estimate specification 9 by regressing assessments on the test score and the same-race dummy. With that approach, the estimate of δ will not be consistent because it will capture part of student ability instead of capturing only teacher biases:[28]

$$plim(\text{Estimator of } \delta) = \delta + \alpha \cdot \lambda\theta \quad (10)$$

where δ is the coefficient of teacher bias, and $\theta = \text{var}(\nu)/[\text{var}(\nu) + \text{var}(Ability)]$ and $\lambda = \text{Cov}(Same\ Race,\ Student\ Ability)/\ \text{Var}(Same\ race)(1 - \text{Corr}(Same\ race,\ Test\ score)^2)$. If, as suggested by Dee (2005), student ability is higher when taught by a same-race teacher, ability and the same-race dummy are positively correlated, $\lambda > 0$, $\alpha \cdot \lambda\theta > 0$ and the effect α of same-race teachers on assessments will be overestimated.[29]

If the relative size θ of the measurement error were known, an unbiased effect of same-race teachers on assessments could be recovered. This unbiased

27. Formally, if the test score is a sufficient statistic for student ability.
28. The algebra is a particular case of the formulas of Greene (2011); plim denotes the probability limit of the estimate.
29. This result is very close to equations of the statistical discrimination literature (see, e.g., Phelps 1972). On the labor market, the employer's hiring decision may depend on the race of the job candidate because the candidate's education, experience, and other covariates are not sufficient statistics for the candidate's productivity.

Table 7. Could Measurement Error in Test Scores Explain the Same-Race Effect?

	Mathematics – Size of Measurement Error in Test Scores						
	θ = 0.00	**θ = 0.05**	**θ = 0.10**	**θ = 0.15**	**θ = 0.20**	**θ = 0.25**	**θ = 0.30**
Same Race	0.711**	0.668**	0.620*	0.566*	0.506*	0.438*	0.360*
	(0.211)	(0.189)	(0.267)	(0.241)	(0.252)	(0.212)	(0.142)
Corrected Test Score	0.241**	0.254**	0.268**	0.284**	0.301**	0.322**	0.345**
	(0.010)	(0.008)	(0.013)	(0.011)	(0.009)	(0.015)	(0.017)
	English – Size of Measurement Error in Test Scores						
	θ = 0.00	θ = 0.05	θ = 0.10	θ = 0.15	θ = 0.20	θ = 0.25	θ = 0.30
Same Race	0.435*	0.384*	0.327**	0.264*	0.193	0.113	0.021
	(0.174)	(0.152)	(0.090)	(0.123)	(0.153)	(0.143)	(0.178)
Corrected Test Score	0.313**	0.330**	0.348**	0.368**	0.391**	0.417**	0.446**
	(0.007)	(0.006)	(0.008)	(0.006)	(0.007)	(0.008)	(0.011)

Notes: Test scores have a standard deviation of 10 and a mean of 50. All regressions are two-way fixed-effects regressions with both a child and a teacher fixed effect. Standard errors are bootstrapped, clustered by student. The corrected test score is such that equation 13 holds.
*Statistically significant at the 5% level; **statistically significant at the 1% level.

estimate of same-race effects is obtained by regressing assessments on a corrected value of the test scores, defined as follows:

$$Corrected\ Test\ score_{i,f,t} = \theta \cdot E[Test\ score_{.,f,t} | Same\ race] + (1-\theta) \cdot Test\ score_{i,f,t}. \quad (11)$$

When we estimate specification 8 replacing the test with this test score, the estimator of the same-race effect will be an unbiased estimate of same-race effect on teacher assessments δ.

This holds if we know the size of measurement error θ. But θ is unknown, and we estimate the parameter of interest δ using different values of θ. The lowest value of measurement error θ that cancels the estimate of the effect of a same-race teacher on assessments yields an estimate of the lowest amount of measurement error that could account for the baseline results. Results for the baseline specifications with corrected test scores are presented in table 7.[30]

For mathematics test scores, a measurement error of more than 30 percent is required to render the coefficient nonsignificant, and additional results show that 40 to 50 percent of measurement error is required to cancel the point estimate. For English, a 20 percent measurement error makes the coefficient nonsignificant, and additional results show that measurement error of

30. Results for measurement error above 30 percent are available upon request.

40 percent cancels the point estimate. In short, a significant amount of measurement error would be necessary to cancel coefficients. Even though this statistic does not exclude the potentially confounding effect of measurement error, it does indicate that only a large amount of measurement error in test scores would alter the conclusions.

Grading on a Curve

Teacher assessments in each subject are an average of ten different assessments on a scale of 1 to 5, which is then standardized to a mean of 50 and a standard deviation of 10. Athough the skills that each assessment evaluates are clearly defined by the survey questionnaire, there is no guideline as such on what should be the standard deviation of assessments across students within a classroom, or what exact proficiency level justifies awarding a 5 or a 4. It may well be that the teacher implicitly ranks students within a classroom.[31]

The implications of grading on a curve for the measurement of a bias in favor of same-race students are multiple. First, teacher assessments may not be directly comparable to test scores, as they will reflect a ranking of students within a classroom, while test scores have a common scale for all participating students. Second, the teacher assessment of a given student will be correlated with peers' average test score in the classroom. Third, if peer group ability is significantly correlated with being assigned a same-race teacher, the estimated OLS effect of a same-race teacher on teacher assessments in specification 1 will be biased.

If teacher assessments reflect a ranking of students within a classroom rather than a measure on a common scale, we should expect black students to get lower assessments than white students. Indeed, consider a simple model where there are only two students in each classroom, and each student can have either a low teacher assessment (a_l) or a high teacher assessment (a_h). A student gets a high assessment if he is the student with the highest ability in the classroom. Student ability is denoted ω, and follows a cumulative distribution function $F(\omega)$. Each student can be either white, $r = w$, or minority, $r = m$. The cumulative distribution function given the student's race r is denoted $F(\omega|r)$. Then a student gets a high assessment a_h if his ability is higher than his peer's ability.

Hence, a student of race r has a high teacher assessment with probability $P(a = a_h|r,\omega) = P(\omega > \acute{\omega}|r,\omega) = F(\acute{\omega}|r,\omega)$. For simplicity, assume that peer ability $\acute{\omega}$ is independent of student ability conditional on race, that is, $F(\acute{\omega}|r,\omega) = F(\acute{\omega}|r)$.[32] In the data we observe that minority students are in classrooms with lower average test scores. Black students are in classrooms that have

31. Grading on a curve is one of the potential grading practices considered by Figlio and Lucas (2004).
32. Similar results hold if students are sorted by ability across classrooms.

an average test score 13.7 percent of a standard deviation below the average test score of white students' peers. We also observe that the distribution of black students' peers' test scores is strictly worse than white students' peers' test scores. Formally, white students' peers' test score distribution first-order stochastically dominates black students' peers' test score distribution, $F(\acute{\omega}|w) < F(\acute{\omega}|b)$.

Then, at a given ability level ω, white students are less likely to get a high assessment than black students:

$$P(a = a_h|w, \omega) - P(a = a_h|b, \omega) = F(\acute{\omega}|w, \omega) - F(\acute{\omega}|b, \omega) < 0.$$

If teacher assessments reflect a ranking in the classroom, we should thus observe that, conditional on test scores, minority students get higher teacher assessments than white students. But results (available from the author) show a nonsignificant or a negative and significant effect of race on teacher assessment conditional on test scores. Another regression suggests a nonsignificant effect of peers' test scores on teacher assessments. Such results make it unlikely that teacher assessments are a ranking of students within each classroom.

The baseline effect of a same-race teacher on teacher assessments of table 4 and specification 8 is also not likely to be affected by teachers grading on a curve within each classroom. Column (1) of table 8 suggests that being assigned a same-race teacher is negatively correlated with peers' test scores. But column (2) of table 8 shows that being assigned a same-race teacher is not significantly correlated with peers' test scores when controlling for a student effect and teacher observables. Column (3) of the same table estimates the same-race effect in mathematics. The novelty compared to baseline specification 8 is that the specification controls for peers' test scores. The estimate (+0.701) is virtually unchanged compared to table 4. Similar results, available from the author, hold in English.

Results with All Racial Interaction Terms

What races drive the results of the main specification? We disentangle the effects of different racial interactions in specification 8, by replacing the Same Race dummy by a set of dummies, one dummy for each interaction between the teacher's and the student's race:

$$\begin{aligned} TA_{i,f,t} = {} & teacher_{i,f,t} + \text{constant} + \alpha TS_{i,f,t} + student_{i,f} \\ & + \sum_{r \neq r'} \delta_{r,r'} Dummy(teacher\ race = r) x Dummy(student\ race = r') \\ & + grade_{t,f} + \varepsilon_{i,f,t} \end{aligned} \tag{12}$$

Table 8. Grading on a Curve Hypothesis

	Mathematics		
	(1)	(2)	(3)
	Peers' Test Scores	**Same Race Teacher**	**Teacher Assessment**
Same Race	–0.609**	–	0.701**
	(0.168)		(0.247)
Peers' Average Test Score	–	–0.002	0.065
		(0.002)	(0.061)
Test Score	–	–0.002**	0.264**
		(0.001)	(0.025)
Student and Teacher Race and Gender	Yes	Yes	No
Student Effects	No	Yes	Yes
Teacher Effects	No	No	Yes
F Statistic	114.5	13.5	4.2
R^2	0.13	0.82	0.79
Observations	48,065	48,065	48,065

Notes: Standard errors clustered by student. Coefficients have similar significance levels when clustering by classroom.
**Statistically significant at the 1% level.

where there is one racial interaction dummy for each pair of races r,r'. Dummy(teacher race = r) × D(student race = r') = 1 if the teacher's race is r and the student's race is r', and 0 otherwise. The effects of interest are the coefficients $\delta_{r,r'}$. The omitted dummy variables are the dummies for a teacher and a student of the same race, hence coefficients are interpreted relative to the assessment given by a same-race teacher.

Results are presented in table 9.[33] In mathematics, being assessed by a white teacher lowers the assessment of Hispanic children by 17.3 percent of a standard deviation, compared with being assigned by a Hispanic teacher (the same-race interaction dummy is omitted). The interaction between white teachers and black students is not significant, but the coefficient's order of magnitude is comparable to baseline estimates. In English, the interaction is significant. White teachers give lower assessments to black children, lower by 11.1 percent of a standard deviation. They also give lower assessments to Hispanic children, by 14.8 percent of a standard deviation.

33. Results from very small minority groups (Pacific Islanders, American Indians) may not be robust. All racial interactions are included in the regressions but only coefficients for blacks, Hispanics, and whites are reported in the table.

Table 9. Effects of All Racial Interactions Terms on Teacher Assessments

	Mathematics Teacher Assessment			English Teacher Assessment		
	(1)			(2)		
	Race of the Student			Race of the Teacher		
Race of the Teacher	**White, non-Hispanic**	**Black**	**Hispanic**	**White, non-Hispanic**	**Black**	**Hispanic**
White, non-Hispanic	Ref.	–0.616	–1.728**	Ref.	–1.110**	–1.480**
		(0.512)	(0.627)		(0.300)	(0.221)
Black	–0.590	Ref.	–1.337	0.530	Ref.	–0.980
	(0.479)		(0.872)	(0.414)		(0.756)
Hispanic, Any Race	0.899	0.371	Ref.	1.684**	–0.643	Ref.
	(0.675)	(1.697)		(0.568)	(0.741)	
Test Score		0.241**			0.314**	
		(0.009)			(0.008)	
F Statistic		4.2			5.6	
R^2		0.787			0.774	
Student Effects		Yes			Yes	
Teacher Effects		Yes			Yes	
Grade Effects		Yes			Yes	
Observations		48,065			67,855	

Notes: This table presents the results of two separate regressions, each with the full set of interactions between the teacher's race and the child's race. Only the three largest minority group interactions are displayed in this table, but other interactions are included in the regressions. Ref. = interaction dummy omitted from the regression.
**Statistically significant at the 1% level.

Despite the size of standard errors, statistical tests show that black teachers give significantly higher English assessments to white students than white teachers to black students. Hispanic teachers, too, tend to give higher assessments in English to white students than white teachers to Hispanic students.[34] In mathematics, white teachers give significantly lower assessments to Hispanic students than to white and black students.[35]

Table 9 also shows that Hispanic teachers tend to give higher grades to white students than to Hispanic students in English. Hence most of the

34. A post-regression χ^2 test rejects the equality of coefficients "white teacher–black student" and "black teacher–white student," as well as the equality of coefficients "white teacher–Hispanic student" and "Hispanic teacher–white student." The χ^2 statistic is 15.28 (respectively, 15.11) with a *p*-value of 0.0001 (respectively, 0.0001).
35. The "white teacher–Hispanic student" coefficient is significant. Moreover, a χ^2 test rejects the equality of the "white teacher–Hispanic student" coefficient and the "white teacher–black student." The statistic equals 4.62 and the *p*-value is 0.0316.

same-race effect on teacher assessments is driven by the behavior of white teachers toward black and Hispanic students.

Policy Implications

Racial Gaps in Test Scores and in Teacher Assessments

Columns (1) to (4) of table 10 estimate racial gaps in test scores and in teacher assessments from kindergarten to grade 5 for both mathematics and English.[36] As documented in the literature, the gap between white and black test scores increases from kindergarten to grade 5: from 63 percent to 93 percent of a standard deviation in mathematics, and from 45 percent to nearly 80 percent of a standard deviation in English.

However, teacher assessments present a different picture. The white–black teacher assessment gap narrows slightly, decreasing from 47 percent to 45.5 percent of a standard deviation in mathematics and from 42 percent to 38.5 percent of a standard deviation in English. It is interesting that, over the same period, the fraction of black students assessed by a same-race teacher increases from 27.3 percent in kindergarten to 34.5 percent in grade 5, and the fraction of white students assessed by a same-race teacher remains relatively constant, at 92 percent.

Because teacher assessments may depend on teachers' identities, columns (9) to (12) present teacher assessment racial gaps while controlling for teachers' race and for teacher–student racial interaction dummies.[37] In these columns, the gap in teacher assessments increases from fall kindergarten to grade 5, from 37 percent to 49 percent of a standard deviation in mathematics, and from 46.6 percent to 49 percent of a standard deviation in English. The racial teacher assessment gap is increasing only when controlling for teachers' race and teacher–student racial interactions.[38]

For Hispanic students, gaps in teacher assessments narrow faster than gaps in test scores. The white–Hispanic test score gap declines from 78 percent to 54 percent of a standard deviation in mathematics (a reduction of 24 percentage points [p.p.]); the white–Hispanic teacher assessment gap declines from 57 percent to 22 percent of a standard deviation in mathematics (a reduction of 35 p.p.). In columns (9) and (10), where regressions incorporate teachers' race dummies and teacher–student racial interaction dummies, the gap in teacher assessment of student mathematics skills goes from 43 percent to 28 percent of a standard deviation (a 15-p.p. reduction). The situation is similar

36. Spring kindergarten, spring grade 1, and spring grade 3 are omitted from the table to save space, but the gaps evolve in the same manner from fall kindergarten to spring grade 5.
37. The full set of variables Dummy(Student race = r) × Dummy(Teacher race = r') for all pairs of races r and r'.
38. Including other teacher observables as controls, such as gender, experience, tenure, and teacher fixed effects, does not affect white–black teacher assessment gaps.

Table 10. Racial Gaps in Test Scores and in Teacher Assessments

	Test Score				Teacher Assessment				Teacher Assessment			
	Mathematics		English		Mathematics		English		Mathematics		English	
	Fall Kindergarten (1)	Spring Grade 5 (2)	Fall Kindergarten (3)	Spring Grade 5 (4)	Fall Kindergarten (5)	Spring Grade 5 (6)	Fall Kindergarten (7)	Spring Grade 5 (8)	Fall Kindergarten (9)	Spring Grade 5 (10)	Fall Kindergarten (11)	Spring Grade 5 (12)
Black	–6.236**	–9.287**	–4.538**	–7.957**	–4.741**	–4.555**	–4.145**	–3.858**	–3.744**	–4.900**	–4.662**	–4.933**
	(0.296)	(0.534)	(0.281)	(0.386)	(0.401)	(0.551)	(0.330)	(0.417)	(1.181)	(0.974)	(0.771)	(0.711)
Hispanic	–7.785**	–5.387**	–5.251**	–6.264**	–5.568**	–2.176**	–4.570**	–2.427**	–4.306**	–2.761**	–4.579**	–3.094**
	(0.309)	(0.421)	(0.271)	(0.303)	(0.346)	(0.430)	(0.289)	(0.317)	(1.139)	(0.886)	(0.744)	(0.634)
Asian	1.350*	0.615	2.357**	–1.374**	–0.378	2.383**	–0.607	1.604**	0.663	1.444	–0.570	0.477
	(0.574)	(0.780)	(0.525)	(0.502)	(0.765)	(0.722)	(0.509)	(0.489)	(1.358)	(1.070)	(0.871)	(0.722)
Teacher Race and Racial Interaction Terms	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes
Observations	11,600	5,233	16,304	10,627	11,600	5,233	16,304	10,627	11,600	5,233	16,304	10,627
R^2	0.12	0.12	0.07	0.11	0.07	0.04	0.05	0.05	0.07	0.04	0.05	0.05
F Statistic	118.3	62.3	89.6	94.4	46.4	15.9	55.3	46.7	32.3	10.9	40.48	33.1

*Statistically significant at the 5% level; **statistically significant at the 1% level.

for assessments of English skills: although the gap in test scores rises by 10 p.p., the gap in teacher assessments goes down by 35 p.p. With controls, in columns (11) and (12), the gap in teacher assessments falls by only 15 p.p.

Broadly speaking, relying solely on teacher assessments may not provide an accurate description of racial gaps from kindergarten to grade 5. Black–white test score gaps in teacher assessments do not increase from kindergarten to grade 5, whereas racial gaps in test scores suggest that African American students are falling behind. Hispanic–white gaps in teacher assessments narrow faster than gaps in test scores, except when controlling for dummies for the teacher's race and teacher–student racial interaction dummies.

Teacher Assessments and Later Test Scores

The paper's main result will be especially important if teacher assessments reflect perceptions that have a causal impact on student performance in mathematics and English. The effect of more favorable assessments is ambiguous as, on the one hand, studies report that more positive treatment and attitudes toward minority students lead to higher achievement (Casteel 1998); on the other hand, in a survey of existing research, Cohen and Steele (2002) describe the potentially negative impacts of "overpraising" and "underchallenging" students (Mueller and Dweck 1998). Importantly, in this paper's data set, students do not see teacher assessments. Therefore, it is unlikely that teachers were trying to please students by being too positive about their English and mathematics abilities.[39]

Estimating the impact of teacher perceptions on student performance is difficult because a causal estimation requires an experimental setting in which teachers get randomized information on students; typical experiments deceive teachers, inducing them to think more positively about a random subset of students (Jussim and Harber 2005). Experiments are typically performed on relatively smaller samples that are not nationally representative. In the well-known Pygmalion study, a random fraction of students was labeled as bloomers and the impact of this information on students' IQ progress was found significant (Rosenthal and Jacobson 1968). Effects of teacher perceptions on later achievement are still debated (Jussim and Harber 2005).

The challenge with my observational data set is to identify the impact of teacher assessments separately from the impact of teacher quality, which may be correlated with assessments, and from the impact of student ability, which is likely positively correlated with teacher assessments conditional on test

39. My results that white teachers give lower assessments to blacks and Hispanics suggests that teachers were not trying to provide socially desirable answers. Bertrand and Mullainathan (2001) describe such "social desirability" bias in surveys but here a social desirability bias would mean even lower teacher assessments for black and Hispanic students.

scores. Because the data set follows students over time, and because teacher identifiers are available, we can estimate the impact of previous assessments on later scores conditional on student and teacher effects. A student effect controls for student unobservables that do not vary across grades, while the teacher effect controls for teacher quality and other teacher characteristics that affect later test scores:

$$TS_{i,f,t} = \text{constant} + b \cdot TA_{i,f,t-1} + c \cdot TS_{i,f,t-1} + Student_{i,f} + Grade_{t,f} + Teacher_{i,f,t} + Residual_{i,f,t} \quad (13)$$

where notations are as above, $TS_{i,f,t}$ is the test score of child i in field f in grade t, $TA_{i,f,t-1}$ is the subjective assessment of student i in the previous grade, $TS_{i,f,t-1}$ is the test score in the same subject in the previous period, $Student_{i,f}$ is a student effect, $Grade_{t,f}$ is a grade effect, and $Teacher_{i,f,t}$ is a teacher effect.

The coefficient of interest here is b, the effect of the previous teacher assessment on the test score. In such a regression, estimates of the coefficients may be biased due to regression to the mean (Arellano and Bond 1991): A child who has a test score much above the average in, say, grade 1, is likely to have a test score closer to the average in the next period, in grade 3. This typically leads to biases in the estimation of the coefficients of interest b and c (Nickell 1981). To alleviate this issue, the test score $TS_{i,f,t-1}$ is instrumented by test scores from previous grades as in Arellano and Bond (1991) as long as a student effect is included, in columns (2) to (4) and (6) to (8) of table 11. This table shows that, in such specifications, teacher assessments have an effect on later test scores, over and above prior test scores, child fixed effects, and teacher fixed effects. This effect is robust to a variety of specifications with or without the Arellano and Bond (1991) instrument, with or without child and teacher fixed effects, and with or without controls for peers' test scores. A one standard deviation increase in prior teacher assessment is correlated with a 3.7 percent to 8 percent standard deviation increase in next grade's test score, conditional on the effects and the maintained controls.

In the regression, teacher assessments have a greater impact than test scores on later test scores.[40] Also, keeping in mind the limitations of the regression (absence of an experimental design), the results suggest that having a same-race teacher from kindergarten to grade 5 raises teacher assessments by 7 percent of a standard deviation in mathematics (table 4), which raises grade 5 scores cumulatively over the five waves by 2.8 percent of a standard deviation in mathematics. Although only 2.57 percent of white students never

40. But interestingly, results available on request suggest that teacher assessments do not have an impact on test scores *in the same grade*. Teacher assessments have an impact on later test scores but not a significant impact on current test scores.

Table 11. Impact of Teacher Assessments on Later Test Scores

	Mathematics Test Score				English Test Score			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Test Score in Previous Wave	0.779**	0.057**	0.740**	–0.010	0.685**	0.057**	0.655**	0.063**
	(0.004)	(0.011)	(0.005)	(0.012)	(0.004)	(0.006)	(0.004)	(0.007)
Teacher Assessment in Previous Wave	0.100**	0.061**	0.140**	0.080**	0.138**	0.019**	0.168**	0.037**
	(0.004)	(0.007)	(0.006)	(0.013)	(0.004)	(0.005)	(0.004)	(0.007)
F Statistic	10,188.3	30.2	7,288.2	7.3	14,124.5	34.6	11,955.5	7.4
R^2	0.698	0.916	0.779	0.956	0.614	0.827	0.688	0.871
Student Race and Gender	Yes	No	Yes	No	Yes	No	Yes	No
Teacher Race and Gender	Yes	Yes	No	No	Yes	Yes	No	No
Grade Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student Effects	No	Yes	No	Yes	No	Yes	No	Yes
Teacher Effects	No	No	Yes	Yes	No	No	Yes	Yes
Observations	11,103				31,649			

Notes: Standard errors clustered by student. Clustering by classroom yields similar estimates.
**Statistically significant at the 1% level.

have a same-race teacher from kindergarten to grade 5, 54.3 percent of black students and 63 percent of Hispanic students have not had a single same-race teacher during the same period.

5. CONCLUSION

The paper presents evidence that teachers give better assessments to students of their own race, even when controlling for test scores, student unobservables, teacher unobservables, and behavioral measures. Results are not significantly explained by measurement error in test scores or grading on a curve within each classroom. The same-race effect appears as soon as in kindergarten for skills covered by the tests.

The presence of continuous detailed teacher assessments of similar skills as test scores, the longitudinal nature of the data set, and the use of econometric techniques controlling for a large number of teacher and student fixed effects are key ingredients for obtaining this paper's results.

Such evidence of better perceptions of same-race students' performance using national representative data from the early years, with detailed robustness checks, should contribute to the debate in at least two ways. First, shifting from standardized test scores to teacher assessments of students may introduce bias in assessments. Although teachers may have a better grasp of student ability than tests, teachers' perceptions are also affected by race and ethnicity. Second, my results suggest that teachers' perceptions of same-race students explain part of the positive impact of same-race teachers on student test scores, as documented by Dee (2005).

I would like to thank Brian Jacob, Francis Kramarz, Eric Maurin, Jesse Rothstein, Cecilia Rouse, and Timothy Van Zandt, as well as two anonymous referees, for particularly helpful suggestions on previous versions of this paper. I also thank audiences at the London School of Economics, the University of Amsterdam, Uppsala University, and the Industrial Relations Section at Princeton University. I am indebted to Cecilia Rouse for access to the data set. This project was undertaken while visiting Princeton University. For computing and financial support I thank INSEAD, CREST, the London School of Economics, and the Marie Curie Programme. The usual disclaimers apply.

REFERENCES

Abowd, John M., Robert Creecy, and Francis Kramarz. 2002. Computing person and firm effects using linked longitudinal employer–employee dataset. Unpublished paper, Cornell University.

Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. High wage workers and high wage firms. *Econometrica* 67(2): 251–334. doi:10.1111/1468-0262.00020

Achinstein, Betty, Rodney T. Ogawa, Dena Sexton, and Casia Freitas. 2010. Retaining teachers of color: A pressing problem and a potential strategy for "hard-to-staff" schools. *Review of Educational Research* 80(1): 71–107. doi:10.3102/0034654309355994

Arellano, Manuel, and Stephen Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58(2): 277–297. doi:10.2307/2297968

Baltagi, Badi. 2008. *Econometric analysis of panel data*. Hoboken, NJ: Wiley.

Bertrand, Marianne, and Sendhil Mullainathan. 2001. Do people mean what they say? Implications for subjective survey data. *American Economic Review* 91(2): 67–72. doi:10.1257/aer.91.2.67

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. Measurement error in survey data. In *Handbook of econometrics*. vol. 5, edited by James J. Heckman and Edward Learner, pp. 3705–3843. Amsterdam, The Netherlands: Elsevier.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2011. Robust inference with multiway clustering. *Journal of Business & Economic Statistics* 29(2): 238–249. doi:10.1198/jbes.2010.07136

Carpenter, Jeffrey P., Glenn W. Harrison, and John A. List. 2005. Field experiments in economics: An introduction. In *Research in experimental economics* 10, edited by R. Mark Isaac and Douglas A. Norton, pp. 1–15. Bingley, UK: Emerald Publishing.

Casteel, Clifton A. 1998. Teacher–student interactions and race in integrated classrooms. *Journal of Educational Research* 92(2): 115–120. doi:10.1080/00220679809597583

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2005. Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review* 24(4): 377–392. doi:10.1016/j.econedurev.2004.06.008

Cohen, Geoffrey L., and Claude M. Steele. 2002. A barrier of mistrust: How negative stereotypes affect cross-race mentoring. In *Improving academic achievement: Impact of psychological factors on education*, edited by Joshua Aronson, pp. 305–331. Bingley, UK: Emerald Publishing. doi:10.1016/B978-012064455-1/50018-X

Darling-Hammond, Linda, and Ray Pecheone. 2010. Developing an internationally comparable balanced assessment system that supports high-quality learning. Paper presented at the National Conference on Next Generation Assessment Systems, Center for K-12 Assessment & Performance Management, Washington, DC, March.

Dee, Thomas S. 2004. Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics* 86(1): 195–210. doi:10.1162/003465304323023750

Dee, Thomas S. 2005. A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review* 95(2): 158–165. doi:10.1257/000282805774670446

Ferguson, Ronald F. 2003. Teachers' perceptions and expectations and the black-white test score gap. *Urban Education* 38(4): 460–507. doi:10.1177/0042085903038004006

Figlio, David N., and Maurice E. Lucas. 2004. Do high grading standards affect student performance? *Journal of Public Economics* 88(9): 1815–1834. doi:10.1016/S0047-2727(03)00039-2

Fryer, Jr, Roland G., and Steven D. Levitt. 2004. Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics* 86(2): 447–464. doi:10.1162/003465304323031049

Fryer, Roland G., and Steven D. Levitt. 2006. The black-white test score gap through third grade. *American Law and Economics Review* 8(2): 249–281. doi:10.1093/aler/ahl003

Giuliano, Laura, David I. Levine, and Jonathan Leonard. 2009. Manager race and the race of new hires. *Journal of Labor Economics* 27(4): 589–631.

Giuliano, Laura, David I. Levine, and Jonathan Leonard. 2011. Racial bias in the manager–employee relationship: An analysis of quits, dismissals, and promotions at a large retail firm. *Journal of Human Resources* 46(1): 26–52. doi:10.1353/jhr.2011.0022

Gluszek, Agata, and John F. Dovidio. 2010. The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review* 14(2): 214–237. doi:10.1177/1088868309359288

Greene, William H. 2011. *Econometric analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

Gresham, Frank M., and Stephen N. Elliott. 1990. *Social skills rating system (SSRS)*. Circle Pines, MN: American Guidance Service.

Hinnerich, Björn Tyrefors, Erik Höglin, and Magnus Johannesson. 2011. Are boys discriminated in Swedish high schools? *Economics of Education Review* 30(4): 682–690. doi:10.1016/j.econedurev.2011.02.007

Hughes, Jan, and Oi-man Kwok. 2007. Influence of student–teacher and parent–teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *Journal of Educational Psychology* 99(1): 39–51. doi:10.1037/0022-0663.99.1.39

Ingersoll, Richard M., and Henry May. 2011. Recruitment, retention and the minority teacher shortage. Consortium for Policy Research in Education Research Report No. RR-69.

Jackson, C. Kirabo, and Elias Bruegmann. 2009. Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics* 1(4): 85–108.

Jussim, Lee. 1989. Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology* 57(3): 469–480. doi:10.1037/0022-3514.57.3.469

Jussim, Lee, and Kent D. Harber. 2005. Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review* 9(2): 131–155. doi:10.1207/s15327957pspr0902_3

Kirby, Sheila Nataraj, Mark Berends, and Scott Naftel. 1999. Supply and demand of minority teachers in Texas: Problems and prospects. *Educational Evaluation and Policy Analysis* 21(1): 47–66. doi:10.3102/01623737021001047

Lavy, Victor. 2004. Do gender stereotypes reduce girls' human capital outcomes? Evidence from a natural experiment. NBER Working Paper No. 10678.

Lyons, Anthony, and Yoshihisa Kashima. 2003. How are stereotypes maintained through communication? The influence of stereotype sharedness. *Journal of Personality and Social Psychology* 85(6): 989. doi:10.1037/0022-3514.85.6.989

Marcus, Geoffrey, Susan Gross, and Carol Seefeldt. 1991. Black and white students' perceptions of teacher treatment. *Journal of Educational Research* 84(6): 363–367. doi:10.1080/00220671.1991.9941817

Meier, Kenneth J., Joseph Stewart, Jr., and Robert E. England. 1989. *Race, class, and education: The politics of second-generation discrimination*. Madison, WI: University of Wisconsin Press.

Moulton, Brent R. 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics* 72(2): 334–338. doi:10.2307/2109724

Mueller, Claudia M., and Carol S. Dweck. 1998. Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology* 75(1): 33–52. doi:10.1037/0022-3514.75.1.33

Nickell, Stephen. 1981. Biases in dynamic models with fixed effects. *Econometrica* 49(6): 1417–1426. doi:10.2307/1911408

Phelps, Edmund S. 1972. The statistical theory of racism and sexism. *American Economic Review* 62(4): 659–661.

Price, Joseph, and Justin Wolfers. 2010. Racial discrimination among NBA referees. *Quarterly Journal of Economics* 125(4): 1859–1887. doi:10.1162/qjec.2010.125.4.1859

Rosenthal, Robert, and Lenore Jacobson. 1968. *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development*. New York: Holt, Rinehart & Winston.

Rudner, Lawrence M., and William D. Schafer. 2001. *Reliability: ERIC Digest No. ED458213*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Rutherford, Jr., Robert B., Mary Magee Quinn, and Sarup R. Mathur. 2004. *Handbook of research in emotional and behavioral disorders*. New York: Guilford Publications.

Sherman, Thomas M., and William H. Cormier. 1974. An investigation of the influence of student behavior on teacher behavior. *Journal of Applied Behavior Analysis* 7(1): 11–21. doi:10.1901/jaba.1974.7-11

Stangor, Charles, Gretchen B. Sechrist, and John T. Jost. 2001. Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin* 27(4): 486–496. doi:10.1177/0146167201274009

Tourangeau, Karen, Christine Nord, Thanh Le, Alberto G. Sorongon, and Michelle Najarian. 2009. *Combined user's manual for the ECLS-K eighth-grade and K-8 full sample data files and electronic codebooks*. Alexandria, VA: National Center for Education Statistics.

Van Ewijk, Reyn. 2011. Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review* 30(5): 1045–1058. doi:10.1016/j.econedurev.2011.05.008

Wilson, Robert J., and Rhonda L. Martinussen. 1999. Factors affecting the assessment of student achievement. *Alberta Journal of Educational Research* 45(3): 267–277.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.