Assessed by a Teacher Like Me: Race, Gender and Subjective Evaluations*

Amine Ouazad†

September 2008

Abstract

The underrepresentation of minority teachers and male teachers remains an issue in U.S. elementary education, and there is evidence that racial interactions partly shape student performance. However, there is little work on discrimination within the classroom. Do teachers give better grades to children of their own race, ethnicity, or gender? A U.S. nationally representative longitudinal data set that includes both test scores and teacher assessments offers a unique opportunity to answer this question. I look at the effect of being assessed by a same-race or same-gender teacher conditional on test scores, child effects, and teacher effects. This strategy controls for three confounding effects: (i) children of different races and genders may react differently in the classroom and during examinations, (ii) teachers may capture skills that are not captured by test scores, and (iii) tough teachers may be matched with specific races or genders. Results indicate that teachers give higher assessments to children of their own race, but not significantly higher assessments to children of their own gender. This effect seems to be driven largely by the differential assessments given to non-hispanic black and hispanic children: white teachers give significantly lower assessments to non-hispanic black children and to hispanic children. Results are robust to various checks on endogenous mobility, measurement error, and reverse causality. Moreover, children's behavior is not a significant determinant of same-race or same-gender matching. Finally, relative grading does not explain the main results of this paper.

Keywords: grading, discrimination, stereotype threat, race, gender
JEL Code: C23, I2, J7, J15, J16, J82

*I am indebted to Roland Benabou, Susan Dynarski, Marc Gurgand, Xavier d'Haultfoeuille, Caroline Hoxby, Brian Jacob, Francis Kramarz, Stephen Machin, Eric Maurin, Jesse Rothstein, and Cecilia Rouse for fruitful conversations and comments on preliminary versions of this paper. I also thank the audience of the Labor Lunch Seminar at the Industrial Relations Section, Princeton University, the Lunch Seminar of the Paris School of Economics, the CLOSUP Special Seminar Series at the Ford School of Public Policy at the University of Michigan, and the Labour Markets Seminar at the London School of Economics. This work was undertaken while visiting Princeton University. I thank Professor Cecilia Rouse for access to the restricted-use data set. The author acknowledges financial support from the Economics of Education and Education Policy in Europe network, Princeton University, and CREST-INSEE.

†INSEAD and Centre for Economic Performance at the London School of Economics. INSEAD, Boulevard de Constance, 77300 Fontainebleau, France. amine.ouazad@insead.edu

1 Introduction

Persistent racial and gender gaps are an increasing concern in many countries. In the United States, a typical black 17-year-old reads at the proficiency level of a typical white 13-year-old (Fryer and Levitt, 2006*a*). Girls significantly outperform boys in reading, and boys outperform girls in mathematics. At the macroeconomic level, these gaps may be costly, given that the aggregate return to education is estimated at around 6–10% per year of schooling (Acemoglu and Angrist, 2000). A back-of-the-envelope calculation thus suggests that there could be important gains from reducing the human-capital gap between races and genders.

Of course, those potential gains depend on the cost of reducing racial and gender gaps. Some claim that there are intrinsic differences between races and genders that are not reducible to social or economic factors. One of the most famous arguments is described in Herrnstein and Murray (1994). However, this explanation has been disputed. First, there is no single factor — usually called the g factor — that explains educational or labor market outcomes (Heckman, Stixrud and Urzua, 2006). Second, racial and gender gaps are not constant but rather are increasing with age. Fryer and Levitt (2006*b*) reports that there is no difference in cognitive performance for 1-year-old children. In grade 1, a few covariates for family background are enough to make racial gaps disappear (Fryer and Levitt, 2006*a*). By the end of third grade, covariates do not capture the black–white test score gap (Fryer and Levitt, 2006*a*); and, indeed, the black–white test score gap increases by about 0.1 percent of a standard deviation per year that children are in school. This suggests that teachers' behavior may be a factor.

The explanation may partly rely on the lack of minority teachers in elementary education: the fraction of minority teachers would have to (roughly) double to match the fraction of minority students. In this paper, I look at whether teachers give better subjective assessments to students of their own race and/or gender conditional on test scores. Subjective assessments are pervasive in schools: most teachers fill school records with comments on the child's ability or behavior. Important decisions such as tracking, special education, and ability grouping are partly based on subjective assessments. Moreover, teachers' priors, beliefs, and behavior may be based on what other teachers have reported.

The bottom part of Table I shows how fifth-grade teachers report their grading practices: 11% of white teachers declare they hold all children to the same standards; 19% of non-hispanic black and hispanic teachers provide the same answer. Male teachers too, more often declare holding all children to the same standards: 15% versus 12% for female teachers. Thus teachers' self-reported grading practices vary widely

across race and gender. However, econometric work is needed to reveal teachers' actual grading practices.

I estimate the effect of being assessed by a teacher of the same race on assessments conditional on test scores. I use a unique U.S. longitudinal data set that combines test scores and teacher assessments of children's skills in elementary education. I can thus compare the test scores and teacher assessments when a given child has a same-race teacher versus when the teacher is of a different race. I can also look at differences for a given teacher when assessing same-race children versus children of other races. Combining these two identification strategies, I estimate the effect of same-race and same-gender teaching on assessments conditional on test scores and on child and teacher fixed effects. This addresses three potential identification issues. First, children of different genders and races may behave differently in the classroom and during examinations; examples include differential effect of testing on boys and girls as well as effects arising from stereotype threats (Steele and Aronson, 1998). Second, teacher assessments may capture skills that are not captured by test scores. Third, some teachers may give higher average assessments regardless of their students' race or gender, and this can be correlated with child characteristics.

The data set is the Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999 (ECLS-K), collected by the National Center for Education Statistics of the U.S. Department of Education. It is the first large-scale U.S. study that follows a cohort of children from kindergarten entry to middle school. Hence this is the first paper to look at the discrepancy between test scores and teachers' perception of students' ability that uses a representative longitudinal sample of U.S. children in elementary education. Important findings are that teachers tend to give better assessments to children of their own race and ethnicity, but not significantly higher assessments to children of their own gender. Moreover, this result is mainly due to the lower grades given to non-hispanic black children and to hispanic children.

A number of robustness checks confirm the result of the baseline estimations. I test for endogenous mobility and allow for some correlation between race, gender, and pupil mobility. Furthermore, measurement error checks show that it would take a large amount of measurement error to otherwise explain results. The estimates are also robust to falsification checks in which test scores are regressed on assessments rather than these on teacher assessments. Finally, I show that even if relative ranking and de facto racial segregation could be a potential explanation, controlling for peers' test scores does not change the results.

The analysis of this paper is related to Lavy (2004). Lavy's paper uses high school matriculation exams in Israel. Comparison of blind versus nonblind test scores showed that boys are likely to be overassessed in all subjects. Moreover, the size of the bias was highly sensitive to teachers' characteristics suggesting

that teachers' behavior is causing grade discrimination. This paper differs from Lavy (2004) in at least three ways. First, I compare subjective assessments and test scores, where subjective assessments are based on classroom behavior and coursework; Lavy (2004) compares test scores of blind and nonblind examinations. Second, in Lavy (2004), if tough teachers are more likely to grade boys then the effect of nonblind assessments on boys' test scores could be overestimated. I control for this effect in the ECLS-K by taking into account child and teacher fixed effects.

This paper is also related to a small-scale experiment on fifth-grade teachers in the state of Missouri. Clifford and Walster (1973) sent teachers report cards that included child records randomly matched to photographs, and teachers were asked to assess child ability. The researchers found a significant effect of physical attractiveness on assessments but no effect of gender. Nevertheless, the study raises a number of issues. It is not clear whether this result on Missouri fifth-grade teachers is relevant to assessing discrimination in a representative U.S. classroom, since teachers were assessing students they did not know on the basis of randomly generated school records. The research reported here on the ECLS-K provides a large-scale analysis of teacher assessments in U.S. elementary education.

Better teacher assessments may have beneficial or detrimental effects on performance. On the one hand, better assessments for the same ability level make it easier to get good grades and may therefore decrease the child's marginal benefit of effort (cf. Coate and Loury, 1993). On the other hand, better teacher expectations may raise student expectations or reflect greater investment in the child's education. These stories can be told apart in a controlled experiment. The psychological and educational literature has debated the issue of the effect of teacher expectations at least since the Pygmalion experiment (Rosenthal and Jacobson, 1968). In this experiment, children of an elementary school took a cognitive test at the beginning of the school year. The experimenters then selected 20% of the children and told the teachers that these children were showing "unusual potential for intellectual growth". Empirical results suggested that those labeled as bloomers had significantly higher IQ progress in first and second grade.

Discrimination and the effect of discrimination cannot be jointly identified in the same data set. Identifying discrimination in grading by same-race or same-gender teachers requires a data set such as the ECLS-K, whereas identifying the effect of perceptions requires a controlled experiment. That is an important point, because it is tempting to go further and estimate the effect of perceptions and the amount of discrimination in the same dataset.

Dee (2004) and Dee (2005*b*) show that being taught by a teacher of the same race or gender increases test scores. Empirical results from Project STAR's experiment show that same-race teaching increases test

scores for grade-1 to grade-3 children (Dee, 2004). Other empirical results from the National Education Longitudinal Study show that same-gender teaching increases the test scores of eighth-grade children (Dee, 2005*b*). This paper is different: it estimates the effect of same-race teaching on assessments conditional on test scores. That is, I examine whether teachers have incorrect perceptions of their students' ability — either overestimating or underestimating it. This leads to different policy implications, e.g. including or improving diversity training for teachers.

The rest of the paper is structured as follows. Section 2 presents the Early Childhood Longitudinal Study. It provides a first-hand descriptive analysis of the difference between teacher assessments and test scores as well as some statistics on racial and gender diversity in U.S. elementary education. Section 3 explains main identification issues, the identification strategy, and baseline results. Section 4 checks the robustness of the results. Section 5 shows that assessment rankings are not affected by teacher–pupil racial interactions in the classroom and that relative ranking does not explain the main results. Finally, Section 6 concludes.

2 The Early Childhood Longitudinal Study

In the fall of 1998, the National Center for Education Statistics of the U.S. Department of Education undertook the first national longitudinal study of a representative sample of kindergartners. It started with more than 20,000 children in a thousand participating schools. It then followed children in the spring, in the fall and spring of grade 1, and in the spring of grades 3 and 5. The study's last follow-up will be eighth grade. Follow-ups have combined procedures to reduce costs and maintain the representativeness of the sample. Movers have been randomly subsampled to reduce costs. At the same time, new schools and children have been added to the data set to strengthen the survey's representativeness. In the spring of 1999, part of the schools that had previously declined participation were included. In the spring of grade 1, new children were included; this made the cross-sectional sample representative of first-grade children. All of whom were followed in the spring of grade 3 and 5.

This paper's empirical analysis uses the restricted-use version of the ECLS-K, which contains the race and gender of both teachers and pupils. Observations with missing data on basic variables (test scores, subjective assessments, teachers' and children's race and gender) were deleted. The analysis is done on 48,065 observations in mathematics and 67,085 in English, which is similar to Fryer and Levitt (2006*a*). Weights provided by the survey's designers correct for the subsampling of movers, but most of the analysis

is robust to changes in weights. Race and ethnicity questions for the teacher were combined to match the categories of child's race; hence Hispanic, Any Race is a separate category. "Same race" should therefore be read as "same race (non-hispanic) or both hispanic (any race)"[1].

Test scores were derived from national and state standards, including the National Assessment for Educational Progress (NAEP), the National Council of Teachers of Mathematics, the American Association for the Advancement of Science, and the National Academy of Science. Test scores are based on answers to multiple-choice questionnaires conducted by external assessors. It is a two-stage adaptive test: surveyors administer a routing test and select a longer test of appropriate difficulty. Test scores are made comparable across children using item response theory[2], and items in second-stage forms overlap adjacent forms. Skills covered by the reading assessments from kindergarten to fifth grade include: print familiarity, letter recognition, 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 from kindergarten to fifth grade 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 before the deletion of missing observations.

At approximately the same time, teachers are contacted in their school. Teachers complete one questionnaire per child. Teacher assessments of children's skills, also called the Academic Rating Scale, are separated into three areas: Language and Literacy; General Knowledge; and Mathematical Thinking. I will use English (Language and Literacy) and mathematics assessments (Mathematical Thinking). The instructions make it clear that this is not a test and should not be administered directly to the child. For English and mathematics, teachers answer between seven and nine questions on the childs' proficiency in a set of skills. Answers are on a five-point scale: Not Yet, Beginning, In Progress, Intermediate, Proficient. An overall assessment is computed for each topic. Teacher assessments, like test scores, were standardized to a mean of 50 and a standard deviation of 10 before the deletion of missing observations.

Teachers also report measures of behavior, which are useful as controls. The social rating scale (SRS) has five scales: approaches to learning, self-control, social interaction, impulsive/overactive, and sad/lonely. The Approaches to Learning Scale measures the ease with which children can benefit from

[1]Racial questions follow the 1997 Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity published by the Office for Management and Budget. These standards allow for the possibility of specifying "More than One Race". Nevertheless the share of children who were declared as "More than One Race" is small.

[2]Item response theory computes test scores adjusting for the difficulty of each question. Formally, the probability of a right answer is modelled as $p_i(\theta) = c_i + (1 - c_i)/[1 + \exp(-Da_i(\theta - b_i))]$, where a_i,b_i and c_i are question-specific parameters and θ is child ability.

their learning environment. The Self-Control Scale indicates the child's ability to respect the property of others, control temper, accept peer ideas for group activities, and respond appropriately to pressure from peers. The five Interpersonal Skills items rate 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. Externalizing Problem Behaviors include acting-out behaviors; the Internalizing Problem Behavior Scale asks about the apparent presence of anxiety, loneliness, low self-esteem or sadness.

Basic children's characteristics are summarized in Table I. The sample is balanced in terms of gender and race. Some racial groups are overrepresented to increase the precision of statistics for subgroups. Moreover, test scores and teacher assessments were standardized to a mean of 50 and a standard deviation of 10 before the exclusion of missing data. This makes test scores and teacher assessments comparable to those in the overall population.

What does it mean to be matched to a teacher of the same race or the same gender? Most children are taught by female white teachers, so the potential advantages of same-race or same-gender teaching will mostly be felt by female or white children. Tables I and II show that only 4.4% of teachers are male, and 47.7% of children are matched with a teacher of the same gender. However, the fraction of male teachers increases over time. Only 2.2% of fall kindergarten teachers are male, but this figure jumps to 15.1% among grade-five English teachers and to 17.4% among grade-five mathematics teachers.

Teachers are also mostly white; Table II reveals that 73.9% of teachers are non-hispanic white in fall kindergarten. This fraction decreases over time before rising in grade 5. Most minority teachers are either hispanic (of any race) or black (African Americans). They predominantly teach to minority children, with their classrooms averaging 81.4% minority children. Column 1 of Table IX shows the regression of a "same race" dummy on pupil characteristics: boys are not significantly more likely to be taught by a teacher of the same race, whereas minority children are systematically less likely to be taught by a teacher of the same race. Non-hispanic black and hispanic children are about 66% less likely to be taught by a teacher of the same race, and this unlikelihood rises to 83% for Asian children.

A first taste of the forthcoming results is shown in the descriptive statistics of Table III. Let's start with mathematics. The difference between test scores and teacher assessments is higher for teachers of the same race as African American children (7% of a standard deviation), hispanic, any race children (26.3% of a standard deviation). These differences are significant at 1%. In English, too, these differences are higher for children matched to a teacher of the same race: 15.8% of a standard deviation for African

American children, 35.3% for hispanic children. The difference is slightly negative for white children in English, but this effect disappears when confounding effects are controlled for. Small minority groups are omitted from these initial tables but they are fully included in all subsequent regressions.[3]

Descriptive statistics for same-gender vs opposite-gender pairings do not display the same clear-cut figures, although the difference between teacher assessments and test scores is lower for girls when matched to a teacher of the same gender. These statistics should not be viewed as causal because they do not control for potentially confounding effects. I describe these in the next section.

3 Identification and Results

3.1 Identification of Teacher Discrimination

Descriptive statistics suggest that, in most minority groups, the teacher assessment–test score gap is higher when student and teacher are of the same race (Table III). This should not be interpreted as a causal effect for a number of reasons.

First, teachers may report skills that are not captured by test scores. The description of the data set makes it clear that, in principle, teacher assessments and test scores cover the same skills. But questions and answers give some leeway. Questionnaires do not formally define the meanings of the five answer categories (Not Yet, Beginning, In Progress, Intermediate, Proficient). Second, boys and girls, white and minority children may display skills differently in the classroom and in a multiple-choice questionnaire. Studies have shown that, for instance, boys react differently to high-stake examinations. Third, some teachers give (on average) higher grades than other teachers for children of the same abilities. The teacher's tendency to give higher grades may be correlated with being of the same race or gender as the students, in which case the gap between test scores and assessments varies with same-race or same-gender teaching without reflecting discrimination.

The baseline specification will attempt to cope with these three potential issues; in this specification, teacher assessments depend on test scores, teacher fixed effects, child fixed effects and a variable indicating whether the child is matched to a teacher of the same race or gender. Formally:

[3]Results available on request.

$$a_{i,f,t} = \mu_{J(i,f,t)} + \delta y_{i,f,t} + u_{i,f} + \alpha_r \text{Same Race}_{i,f,t} + \varepsilon_{i,f,t}, \quad (1)$$

$$a_{i,f,t} = \mu_{J(i,f,t)} + \delta y_{i,f,t} + u_{i,f} + \alpha_g \text{Same Gender}_{i,f,t} + \tau_{t,f} + \varepsilon_{i,f,t}. \quad (2)$$

Here $a_{i,f,t}$ is the teacher assessment of child i in field f (English or mathematics), in period t, whichs runs from fall kindergarten to spring grade 5. The $y_{i,f,t}$ term is the test score, $u_{i,f}$ is the child effect of child i in field f, and $\mu_{J(i,f,t)}$ is the teacher effect. Same Race$_{i,f,t}$ (resp. Same Gender$_{i,f,t}$) takes value 1 when matched with a teacher of the same race (resp. gender) and 0 otherwise. $\tau_{t,f}$ is a season & year-specific fixed effect.

$u_{i,f}$ captures non–time–varying individual characteristics that may have an effect on assessments regardless of the teacher. For instance, this term may capture behavior, which teachers typically include in their assessments. Boys may react differently to classroom exercises, which are assessed by the teacher, and to the multiple-choice questions of the ECLS-K.

The inclusion of teacher effects $\mu_{J(i,f,t)}$ attempts to cope with the third identification issue. If the teacher's grading practice $\mu_{J(i,f,t)}$ is correlated to same-race or same-gender teaching, then the OLS estimates of α_g and α_r might be biased. Hence teacher effect $\mu_{J(i,f,t)}$ is used to capture these permanent average differences between teachers.[4]

The model is estimated using a preconditioned conjugate gradient method described in Abowd, Creecy and Kramarz (2002).[5] All estimations have converged with a numerical precision of 10^{-15}. Bootstrap was used to compute standard errors, as described in Efron and Tibshirani (1994); specifically, block bootstrap was performed (i.e. simple random sampling of children, which takes into account the correlation of residuals across observations of the same child).

As in Abowd, Kramarz and Margolis (1999) and Kramarz, Machin and Ouazad (2007), children moving from/to a same-race teacher identify the effect of same-race assessments conditional on test scores. The identification of specifications (1) and (2) therefore requires sufficient and exogenous mobility[6].

Exogenous mobility is best understood when comparing the progress of a child in terms of assessments

[4] Another identification issue may arise if some teachers "spread" their assessments more than others, in which case δ may vary from teacher to teacher. However, estimations are too imprecise when allowing this flexibility. Results are available on request.

[5] I have developed a set of STATA packages available on the web by typing 'ssc install a2reg' on the command line.

[6] Sufficient mobility can be properly defined. As in Abowd et al. (1999) and Kramarz et al. (2007), two teachers are said to be *connected* when they have taught the same child in different years. This defines a network of teachers connected together through children. For all teachers in the same connex component of the mobility graph, it is then possible to identify teachers' relative toughness in grading.

to progress in terms of test scores. Let's therefore take the first difference of specifications (1) and (2):

$$\Delta a_{i,f,t} = \Delta\mu_{J(i,f,t)} + \delta\Delta y_{i,f,t} + \alpha_r \Delta\text{Same Race}_{i,f,t} + \Delta\varepsilon_{i,f,t}, \quad (3)$$

$$\Delta a_{i,f,t} = \Delta\mu_{J(i,f,t)} + \delta\Delta y_{i,f,t} + \alpha_g \Delta\text{Same Gender}_{i,f,t} + \Delta\varepsilon_{i,f,t}. \quad (4)$$

The effect of same-race and same-gender assessments is identified whenever $\Delta\text{Same Race}_{i,f,t}$ and $\Delta\text{Same Gender}_{i,f,t}$ are not correlated with unobserved characteristics that have an impact on the progress in assessments, conditional on the variation in teacher effects $\Delta\mu_{J(i,f,t)}$ and the progress in test scores $\Delta y_{i,f,t}$. In other words, child mobility should not be driven by unobserved time-varying shocks that affect teacher assessments conditional on the other covariates. Section 4.1 suggests that this issue is not affecting the empirical results.

3.2 Baseline Results

Baseline results suggest that teachers indeed give better assessments to pupils of their race, but not significantly better assessments to pupils of their own gender. The effect is sizeable: it is between 1/10 and 1/5 of the black–white teacher assessment gap and about 1/3 of the hispanic–non-hispanic teacher assessment gap.

Baseline results are presented in Table IV. OLS estimates indicate that children who are assessed by same-race teachers also have higher math assessments around 2.8% of a standard deviation higher. However, this is not likely to be the causal effect of same race assessments for reasons outlined previously. Column 2 gives the estimate when controlling for child effects. This estimate is higher than the baseline OLS, which suggests that the child fixed effect is negatively correlated with same-race pairings. Most teachers are female non-hispanic white, therefore either on average all teachers give lower assessments to white children or white children respond differently when in the classroom than when facing an assessor.

Column 3 gives the estimate when controlling for teacher fixed effects. Again, the estimate is higher than the OLS estimate of column 1, which implies that the teacher fixed effect is negatively correlated with same-race pairings. Teachers who give lower assessments are matched with children of the same race. Again, a majority of teachers are white females, and a possible story is that these teachers are tougher than teachers of other races.

Finally, column 4 gives the estimate when controlling for both children and teacher fixed effects. The

estimate is similar to the estimates of columns 2 and 3. Column 4 is my preferred estimate for the effect of same-race matching on assessments conditional on test scores. It indeed addresses the three important identification issues already described. On average, children who are assessed by a teacher of the same race have a higher mathematics assessment — around 7% of a standard deviation higher.

Turning to English assessments, the OLS estimate and the child fixed effects are roughly similar: children who are assessed by a teacher of the same race also have a higher English assessment, by around 4% of a standard deviation. Column 7 shows that controlling for teacher fixed effects actually increases the estimate, suggesting the same correlation between grading practices and same-race matching as for math assessments. Column 8 shows the estimate when controlling for both children and teacher fixed effects. Surprisingly, the effect is of the same magnitude as the OLS and the child fixed effect estimates; this is due to the negative correlation between child and teacher fixed effects. Results indicate that being matched with a teacher of the same race increases assessments by around 4% of a standard deviation, conditional on test scores and on children and teacher fixed effects.

The gender and racial gaps in teacher assessments are shown in Table V. This table is useful for comparing the gaps in assessments to the magnitude of the effect. In mathematics, the effect of same-race assessments is around 7% of a standard deviation, approximately 1/3 of the black–white teacher assessment gap and 1/5 of the hispanic–non-hispanic teacher assessment gap. In English the effect of same-race assessments is around 4.1% of a standard deviation, which explains about 1/10 of the black–white teacher assessment gap. Overall, the effect of race interactions on assessments accounts for between 1/10 and 1/3 of the teacher assessment gaps.

3.3 Analysis of Child and Teacher Effects

Child fixed effects are interpreted as: (i) differential behavior in the classroom and during tests; (ii) unobserved characteristics that teachers may include in their assessment; (iii) average grading discrimination. Column 1 of Table VI shows that boys' fixed effects are 19% of a standard deviation lower, controlling for race. Controlling for teacher-reported child's behavior, the difference between boys' and girls' fixed effects is much smaller (7% of a standard deviation). This indicates that teacher assessments partly include the child's behavior. The same reasoning for other rows of columns 1 and 2 of Table VI suggests that lower fixed effects for minority children are partly due to the inclusion of behavior in teachers' assessments.

Teacher grading practices are captured by teacher fixed effects. The fixed effects are higher when teachers give better assessments regardless of the student's race or gender. Columns 3 and 4 of Table

VI show results of the analysis of teacher fixed effects. Male teachers' effects are 5.3% of a standard deviation higher, suggesting that male teachers give better assessments on average. Black, hispanic, and Asian teachers' effects are between 1% and 2.5% of a standard deviation lower. These correlations are stable when controlling for tenure and experience, even though the proportion of minority teachers has steadily declined in the last decades.

3.4 Breaking Down Results by Race

Results have suggested that teachers give higher grades to children of their own race conditional on test scores and children's and teachers' constant characteristics. What races drive these results? In order to disentangle the effects of different racial interactions, I will estimate a specification in which the Same Race dummy is split into multiple dummies, one for each interaction between the teacher's and the student's race. This will allow for heterogeneous effects, race by race. The specification is similar to baseline specification (1).

$$a_{i,f,t} = \mu_{J(i,f,t)} + \delta y_{i,f,t} + \sum_{r \neq r'} \alpha_{r,r'} D(r,r') + u_{i,f} + \varepsilon_{i,f,t}, \tag{5}$$

where r denotes teacher's race and r' student's race. Pupil i's assessment $a_{i,f,t}$ in field f in period t depends on test scores $y_{i,f,t}$, a set of interactions between the teacher's and the student's race $D(r,r')$, child effects $u_{i,f}$ and teacher effects $\mu_{J(i,f,t)}$.

Results are presented in Table VII.[7] This more refined analysis of racial interactions gives a better view of teacher perceptions. In mathematics, being assessed by a white teacher lowers the assessment of hispanic children by 17.3% of a standard deviation. 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% of a standard deviation. They also give lower assessments to hispanic children, by 14.8% of a standard deviation.

Despite the size of standard errors, statistical tests show that black teachers give significantly higher English assessments to white students that white teachers to black students. Hispanic teachers, too, tend

[7] 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 on the table.

to give higher assessments in English to white students than white teachers to hispanic students.[8] In mathematics, white teachers give significantly lower assessments to Hispanic students than to white and black students.[9]

One result from Table VII departs from the idea that same-race assessments result in higher grades: hispanic teachers tend to give higher grades to white students than to hispanic students in English. Overall results broken down by race reveal that the strongest interactions occur between white teachers and black students and between white teachers and hispanic students.

3.5 Do Female Teachers Give Better Assessments to Girls?

It has not been found that teachers give significantly higher grades to children of their own gender conditional on test scores and on children's and teachers' constant characteristics. However, it may be possible that this average effect for both male and female teachers is due to the combination of opposite effects for same-gender teacher–pupil pairings.

I therefore put forward a specification in which heterogenous effects are allowed. In the same way as in Section 3.4,

$$\begin{aligned} a_{i,f,t} &= \mu_{J(i,f,t)} + \delta y_{i,f,t} \\ &\quad + \alpha_{\text{male}}(\text{Male Teacher and Male Pupil})_{i,f,t} \\ &\quad + \alpha_{\text{female}}(\text{Female Teacher and Female Pupil})_{i,f,t} \\ &\quad + u_{i,f} + \varepsilon_{i,f,t} \end{aligned} \tag{6}$$

Pupil i's assessment $a_{i,f,t}$ in field f in period t depends on test scores $y_{i,f,t}$, a $(\text{Male Teacher and Male Pupil})_{i,f,t,\text{Male}}$ dummy, a $(\text{Female Teacher and Female Pupil})_{i,f,t,\text{Female}}$ dummy, child effects u_i, and teacher effects $\mu_{J(i,f,t)}$.

Empirical results (presented in Table VIII) show that male teachers are more likely to give higher assessments to male children in mathematics, increasing them by 6.5% of a standard deviation. Other coefficients are not significant.

[8] 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 (resp. 15.11) with a p-value of 0.0001 (resp. 0.0001).

[9] 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.

4 Discussion

4.1 Are Disruptive Children Assigned to Teachers of Their Own Race?

The baseline model described in equations (1) and (2) is not identified if (i) the child's behavior is implicitly part of teacher assessments and (ii) the child's behavior makes it more likely to be taught by a teacher of the same race. This section shows that there is little correlation between being assigned a teacher of the same race and measures of behavior.

Studies in psychology have shown that family events are correlated with child behavior: children who witness domestic violence suffer from low self-esteem, anxiety, depression and behavior problems (Hughes, 1988); physically abused adolescents have significantly higher prevalence rates of depression, conduct disorder, internalizing and externalizing behavior problems, and social deficits (Pelcovitz, Kaplan, Goldenberg, Mandel, Lehane and Guarrera, 1994). Family events may therefore drive behavioral changes.

Moreover, the economics literature shows that teachers are not randomly assigned to students (Rothstein, 2008). Clotfelter, Ladd and Vigdor (2005) suggests that novice teachers are assigned to classrooms in a way that disadvantages black students. In this analysis, if students who become disruptive are assigned to same-race teachers then I overestimate the effect of same-race assessments. A good test is therefore to regress probabilities of being matched to a teacher of the same race on changes of behavior.

Columns 3 and 4 of Table IX show that there is no significant effect of behavior on the probability of being matched with a teacher of the same race.[10] There is little correlation between behavior and same-race teaching when controlling for teacher fixed effects, and it disappears altogether when controlling for both child fixed effects and teacher fixed effects. This table is for mathematics, and a similar table is available for English teachers. The ECLS-K contains multiple measures of behavior reported by the teacher and by the parents. Table IX uses teacher-reported measures of behavior because they are likely to be more relevant than parental measures in the teacher assignment process.[11]

4.2 Can Measurement Error Explain the Results?

Test scores of multiple-choice questionnaires are usually noisy measures of underlying ability (Rudner and Schafer, 2001). Random error may be introduced in the design of the questionnaire; distractors (wrong options) may not be effective, or may be partially correct; and items may be either difficult or not difficult

[10] This is a linear probability model. Conditional logits allow for the estimation of discrete models with controls for unobservable heterogeneity, and their estimation yields similar results. Conditional logits do not allow the introduction of both student and teacher unobserved heterogeneity.

[11] Parental measures could be used, with no significant effect on the findings. Results are available on request.

enough. Measurement error may be also be due to children's behavior, such as sleep patterns, illness, careless errors when filling out the questionnaire, and misinterpretation of test instructions.

Measurement error in test scores could cause bias in my estimation of the effect of same-race/same-gender teachers on assessments. More precisely, most teachers are (non-hispanic) white, and most minority teachers are either hispanic or African American. The Same Race variable will therefore be correlated with the gap between white and black and between white and hispanic children. This means that the effect of same-race assessments could be overestimated. It is therefore important to check whether measurement error could be a potential story for a significant effect of same-race teachers on assessments in Table IV.

A first hint that measurement error may be an explanation comes from the second row of Table IV. The coefficient of test scores in all regressions is lower than 1, although one would naturally expect this coefficient to be equal to 1, given that both assessments and test scores have a standard deviation of 10. But constraining this coefficient to be equal to 1 does not significantly alter the coefficients of interest (first row of Table IV)[12].

So what measurement error can explain the baseline estimates? Assume that test scores are noisy measures of the child's underlying ability:

$$y_{i,f,t} = y^*_{i,f,t} + \nu_{i,t}. \tag{7}$$

I assume that measurement error is classical (i.e. $\nu_{i,t}$ is not correlated with ability). In other words, (7) assumes that ability is as precisely measured for low-performing children, average children, and high-performing children.

For the sake of clarity, I drop fixed effects in the so-called structural equation:

$$a_{i,f,t} = \mu + \delta y^*_{i,f,t} + \alpha_r \text{Same Race}_{i,f,t} + \varepsilon_{i,f,t}, \tag{8}$$

where teachers' assessments are based on true ability $y^*_{i,f,t}$ rather than test scores $y_{i,f,t}$. The econometrician does not observe $y^*_{i,f,t}$ and so must estimate equation (8) by regressing on $y_{i,f,t}$. Then, both the estimate of δ and the estimate of α_r will be biased:

$$\hat{\alpha}_{r,\text{OLS}} = \alpha_r + \delta \cdot \lambda\theta, \tag{9}$$

where

[12] Results available on request.

$$\theta = \text{var}(\nu)/[\text{var}(\nu) + \text{var}(y^*)] \quad \text{and} \tag{10}$$

$$\lambda = \frac{\text{cov}(\text{Same Race}, y^*)}{\text{var}(\text{Same Race})(1 - \text{corr}(\text{Same Race}, y^*)^2)}. \tag{11}$$

Here θ is the size of the measurement error. If, as suggested, Same Race and test scores y are positively correlated, then $\lambda > 0$ and the effect α of same-race teachers on assessments will be overestimated. This result is in the same spirit as developments from the literature on measurement error and statistical discrimination (see e.g. Phelps, 1972).

Given the relative size θ of the measurement error, one could estimate the unbiased effect of same-race teachers on assessments. Indeed, the corrected value of the test score may be built as follows:

$$\tilde{y}_{i,f,t} = \theta \cdot E[y_{\cdot,f,t} \mid \text{Same Race}] + (1 - \theta) \cdot y_{i,f,t}. \tag{12}$$

The estimation of specification (1) on the corrected test score $\tilde{y}$ will then give an unbiased estimate of the effect α of same-race teachers on assessments, conditional on test scores.

But the size of measurement error is unknown, so I estimate the parameter of interest α using different values of θ. The lowest size of the measurement error will give an estimate of the measurement error that is required to explain our results.

Results for the baseline specifications with corrected test scores are presented in Table X. For mathematics test scores, a measurement error of 30% is required to make the coefficient nonsignificant. Between 40 and 50% of measurement error is required to cancel the point estimate. For English, the required amount of measurement error is smaller. 10% makes the coefficient nonsignificant, and a measurement error of about 30 to 40% cancels the point estimate. In short, in maths, a significant amount of measurement error would be necessary to cancel coefficients. Even though this statistic does not exclude a potential confounding effect of measurement error, it suggests that only a large amount of measurement error would alter our conclusions.

4.3 Are Teacher Assessment–Test Score Gaps Correlated Across Topics?

So far, the analysis has been carried out separately for English and mathematics. It could be fruitful, though, to investigate whether teachers' perceptions are correlated across topics. More precisely: Are the differences between test scores and teacher assessments correlated in English and in mathematics? On the

one hand, if the gap between assessments and test scores reflects teachers' perceptions, then these gaps should be correlated across topics. From kindergarten to third grade, it is indeed the same teacher who fills out teacher assessment forms in both English and mathematics. On the other hand, if the difference between teacher assessments and test scores reflects only measurement error, then their correlation across topics should be low.

Defining the gaps between assessments and test scores:

$$\begin{aligned} \Delta_{i,\text{Mathematics},t} &= a_{i,\text{Mathematics},t} - y_{i,\text{Mathematics},t} \ , \\ \Delta_{i,\text{English},t} &= a_{i,\text{English},t} - y_{i,\text{English},t} \ . \end{aligned}$$

Table XI shows the correlation of teacher assessment–test score gaps across fields, race by race and gender by gender. We remark that the correlation is significant and above .5 for all races expect Pacific Islanders. Moreover, the correlation is remarkably stable across races (ranging from .445 to .552), indicating that teachers' perceptions are correlated across fields regardless of race and gender. These figures also suggest that random noise is not likely to explain the main results of this paper.

4.4 Stereotype Threats

Another — important — identification issue arises because students may truly perform better in the classroom when matched to a teacher of the same race. In this case, it is likely that behavior in the classroom will be affected by same-race teaching. There is evidence that stereotype threats can impair both academic performance and psychological engagement with academics (Aronson, Fried and Good, 2002). Wheeler and Petty (2001) reviewed literature on the link between stereotype activation and behavior. Hence five regressions were performed as a test for stereotype threat:

$$b^k_{i,f,t} = m_{J(i,f,t)} + d \cdot y_{i,f,t} + \theta_{i,f} + a_r \cdot \text{Same Race}_{i,f,t} + e_{i,f,t}, \tag{13}$$

where $b^k_{i,f,t}$ is the kth behavioral measure of pupil i in field f in period t (other notation is as before). The interpretation of fixed effects is slightly different than in the previous sections, though. Here, $b^k_{i,f,t}$ is reported by the teacher and the teacher effect $m_{J(i,f,t)}$ is seen as the average behavioral assessment of teacher $J(i,f,t)$. The $\theta_{i,f}$ term is the pupil's average difference between cognitive performance and behavior.

Results are reported in Table XII. There is no significant effect of same-race assessment on behavior,

conditional on test scores in any of the four behavioral dimensions.[13] This suggests that the child's behavior is not significantly affected by same-race teaching conditional on test scores. Stereotype threats are therefore not likely to explain the main results of this paper. These results do not, however, rule out an unconditional effect of same-race teaching on behavior (as in Dee, 2005*a*).

4.5 Increase in Same-Race Teachers for Minority Children Along the Curriculum

There are many more minority teachers in grade 5 than in kindergarten. As a result, minority children are increasingly likely to move from a teacher of another race to a teacher of their own race as they move from kindergarten to grade 5. Moving from/to a teacher of the same race will therefore be correlated with the child's race. This is a potential identification issue in specifications 1 and 2. I will therefore design a specification that allows for some correlation between race, gender, and mobility patterns.

Table XIII shows the average characteristics of children who experience different mobility patterns. '00100' means that the child had a teacher of the same race in spring grade 1 and a teacher of a different race in the other four periods (fall kindergarten, spring kindergarten, spring grade 3 and spring grade 5). Mobility is strongly correlated with race and ethnicity. Only 4% of white children have never been taught by a teacher of the same race, whereas 25% of African American children have always been taught by a teacher of a different race. Column 1 of Table IX shows that although gender is not correlated with same-race teaching, minority pupils are less likely to be matched with a teacher of their own race in the early years of elementary education. There is indeed a correlation between race and mobility patterns, which is not controlled for in specification 1.[14]

It is possible to condition on the whole history of teacher-student matchings as in Card and Sullivan (1988), which inspired Table XIII. Here, I perform a simpler test, introducing child and teacher fixed effects in the first differenced equation. This allows for some correlation between mobility and children's observed and unobserved characteristics.

$$\Delta a_{i,f,t} = \delta \Delta y_{i,f,t} + \alpha_r \Delta \text{Same Race}_{i,f,t} + u_{i,f} + \mu_{J(i,f,t)} + \nu_{i,f,t}, \quad (14)$$

$$\Delta a_{i,f,t} = \delta \Delta y_{i,f,t} + \alpha_g \Delta \text{Same Gender}_{i,f,t} + u_{i,f} + \mu_{J(i,f,t)} + \nu_{i,f,t}. \quad (15)$$

[13] One behavioral measure, Self-Control and Peers, could not be used as a dependent variable because missing observations would have significantly reduced the sample size.

[14] Take the first-differenced version of baseline specification 1. Taking the difference removes the individual fixed effect, which includes race and gender. The specification is therefore not identified if race is correlated with Same $\text{Race}_{i,f,t+1}$ − Same $\text{Race}_{i,f,t}$, which is nonzero when the student moves from/to a same-race to a different-race teacher.

Notation is as before; $u_{i,f}$ is a child fixed effect, and $\mu_{J(i,f,t)}$ is a teacher fixed effect. These two specifications may then account for the observed correlation between race, included in $u_{i,f}$, and mobility patterns ΔSame Race$_{i,f,t}$ and ΔSame Gender$_{i,f,t}$. However, a major disadvantage of this specification, is the increased standard errors that it generates.

Table XIV show the results for specifications (14) and (15). A striking fact is that, although standard errors are wider, point estimates are remarkably similar to the estimates of specifications (1) and (2). Columns 4 and 8 show the estimates for same race pairings on English and math assessments. The effect is not significant for mathematics; it is similar to the baseline estimates (7% in column 4 of Table (IV), and 8% in column 4 of Table XIV). The estimate for English assessments is both significant and close to the baseline estimate: children paired with a teacher of the same race have an assessment that is 4% of a standard deviation higher than other children.

Overall, mobility based on constant observed and unobserved characteristics such as ability, race or gender does not seem to affect baseline estimates.

4.6 Reverse Causality: Do Teacher Assessments Have An Effect on Test Scores?

Baseline results suggest that teachers give significantly higher assessments to children of their own race. However other stories could explain this result. Teacher assessments may be driving test scores as in the Pygmalion experiment (Rosenthal and Jacobson, 1968), such that expectations do affect educational outcomes. In this scenario, the effect of same-race teaching goes from teacher assessments to test scores, and not vice versa. The following specifications test for potential reverse causality, and empirical results suggest that these stories are not relevant.

Hence, in this falsification test, test scores and teacher assessments are reversed: assessments explain test scores rather than the other way around.

$$y_{i,f,t} = \mu_{J(i,f,t)} + \delta a_{i,f,t} + u_{i,f} + \alpha_r \text{Same Race}_{i,f,t} + \varepsilon_{i,f,t}, \quad (16)$$

$$y_{i,f,t} = \mu_{J(i,f,t)} + \delta a_{i,f,t} + u_{i,f} + \alpha_g \text{Same Gender}_{i,f,t} + \varepsilon_{i,f,t}. \quad (17)$$

Notation is as in the baseline specifications (1) and (2). Results are presented in Table XV: Although the OLS estimates are significantly negative, the effect of same-race teaching becomes nonsignificant when adding a child fixed effect in both the mathematics and English specifications. The effect is also

nonsignificant when controlling for both child and teacher fixed effects. This suggests that reverse causality is unlikely to be a viable alternative story.[15]

5 How Do Teachers Order Assessments?

Results suggest that teachers give higher assessments to children of their own race. Are assessments still ranked the same way as test scores? Even if the absolute value of teacher assessments is biased, the ranking of teacher assessments in the classroom might still reflect the ranking of children's cognitive skills.

5.1 Relative versus Absolute Grading

I computed the child's rank in test scores and teacher assessments within surveyed children in the classroom. Teachers fill out assessment questionnaires only for surveyed children; when no discrimination occurs, the ordering of teacher assessments should be similar to the ordering of test scores.

In the econometric specification, the rank in teacher assessments depends on the rank in test scores, a teacher fixed effect, and a child fixed effect as well as a variable indicating whether the teacher is of the same race or gender:

$$\text{Rank in } a_{i,f,t} = \mu_{J(i,f,t)} + \delta \text{Rank in } y_{i,f,t} + u_{i,f} + \alpha_r \text{Same Race}_{i,f,t} + \varepsilon_{i,f,t}, \tag{18}$$

$$\text{Rank in } a_{i,f,t} = \mu_{J(i,f,t)} + \delta \text{Rank in } y_{i,f,t} + u_{i,f} + \alpha_g \text{Same Gender}_{i,f,t} + \varepsilon_{i,f,t}. \tag{19}$$

Rank in $a_{i,f,t}$ is the rank in teacher assessments within surveyed children of the classroom for child i in field f in period t as before. Rank in $y_{i,f,t}$ is the rank in test scores within surveyed children of the classroom; $\mu_{J(i,f,t)}$ is a teacher effect, and $u_{i,f}$ is a child effect. The coefficients of interest are α_r and α_g.

Results are presented in Table XVI. The OLS estimates of same-race teachers are between 0.09 ranks (mathematics) and 0.119 ranks (English). Controlling for child fixed effects, the teacher race effect falls and remains significant only in English (0.06 ranks). This suggests that some children get better rankings regardless of the teacher's race. Two-way fixed-effects results are not significant in mathematics and English.

[15] This assumes that either $y_{i,f,t} = \mu_{J(i,f,t)} + \delta a_{i,f,t} + u_{i,f} + \alpha_r \text{Same Race}_{i,f,t} + \varepsilon_{i,f,t}$ or $a_{i,f,t} = \mu_{J(i,f,t)} + \delta y_{i,f,t} + u_{i,f} + \alpha_r \text{Same Race}_{i,f,t} + \varepsilon_{i,f,t}$ is the underlying structural equation. But if both equations hold the test doesn't allow us to disentangle the two effects. This is then a simultaneous equation problem, and an instrument for test scores or assessments is needed.

Combining these results with the baseline results: teachers tend to give better *assessments* to children of their race and ethnicity, but they do not seem to alter the *ranking* of students of their race or gender.

5.2 A Small Model of Relative Grading

In fact, relative ranking could potentially explain my main results. I design a small model that explains that identification issue and then test the hypothesis on the dataset. The results do not support ranking as a driving force of my results.

For the sake of clarity, I will design a model that does not capture teacher-student interactions (although it could be extended to include them). It can be extended to teacher-student interactions. Assume that teachers order students on a rigid scale and don't care about the absolute value of the assessments. Blacks could be overassessed whenever (i) they are more likely to be compared with other black kids than with white kids and (ii) black kids have, on average, lower test scores. If black students are more likely to be compared with underachievers when matched to a teacher of the same race than when not, then the effect of same-race assessments might simply reflect ranking and not teachers' perceptions.

I design a small model to explain this effect. Each classroom has two students, who can be either black or white. The teacher assessment of a student is either $a = \overline{a}$ or $a = \underline{a}$ depending on the child's ranking in the test scores in the classroom. Each child can be either black ($r = b$) or white ($r = w$). The overall fraction of white kids in the population is π. I will use primes to designate the child's peer (e.g., the peer's race is r').

The probability of getting a high assessment when black and when the test score is y depends on the distribution of test scores and the de facto segregation pattern.

$$\begin{aligned} P(a = \overline{a} \mid r = b, y) &= P(y > y' \mid r = b, y) \\ &= P(y > y' \mid r = b, y, r' = b)P(r' = b \mid r = b, y) \\ &\quad + P(y > y' \mid r = b, y, r' = w)P(r' = w \mid r = b, y) \\ &= P(y > y' \mid r = b, r' = b)P(r' = b \mid r = b) \\ &\quad + P(y > y' \mid r = b, y, r' = w)P(r' = w \mid r = b) \end{aligned}$$

I assume that there is no correlation between the test score and the probability of being matched to a black pupil. Assuming that there is no correlation between test scores in a classroom (i.e. no peer

effects, which is an assumption that can be relaxed), let's say that the distribution of test scores is $f_b(y)$ for blacks and $f_w(y)$ for whites. Moreover, the segregation pattern can be described by a single number $p = P(r' = b \mid r = b)$ that doesn't change with test scores, $r' \perp y \mid r$. Then

$$P(a = \overline{a} \mid r = b, y) \quad = \quad F_w(y) \cdot (1 - p) + F_b(y) \cdot p$$

and, symmetrically for whites,

$$P(a = \overline{a} \mid r = w, y) \quad = \quad F_w(y) \cdot (1 - p') + F_b(y) \cdot p'$$

where $p' = P(r' = b \mid r = w) = \frac{\pi}{1-\pi}(1 - p)$. This leads to the following effect of race on assessments:

$$\begin{aligned} \delta a(y) \quad &= \quad P(a = \overline{a} \mid r = w, y) - P(a = \overline{a} \mid r = b, y) \\ &= \quad [F_w(y) - F_b(y)](p - p'). \end{aligned}$$

If white children have uniformly better test scores and if there is some degree of de facto segregation, then $F_b(y) > F_w(y)$ for all y and $p > p'$. This leads to lower assessments for white children (i.e. $\delta < 0$).

This makes clear that, even in the absence of any form of teacher misperception, there can be effects of the child's race on teacher assessments. This result relies on the relationship between teacher assessments and classroom composition and is therefore testable.

5.3 Controlling for Peers in the Baseline Equation

My regressions, of course, regress assessments on the *interaction* between the teacher's race and the student's race, and not simply on the latter. However, relative grading might still be a cause of spurious results if the Same Race dummy is correlated with classroom composition. In other words, there is a bias if students who move from a same-race teacher to a teacher of a different race are more likely to move to a classroom with worse peers, conditional on child and teacher fixed effects. In this case, peers' test scores are correlated with the Same Race dummy, which invalidates the causal interpretation of the identification strategy.

I design two falsification tests. First, I regress a same-race teacher dummy on the average test score in the classroom either conditional on either teacher or child fixed effects. Second, I include peers' average test score as a control in the baseline regression.

Table XVII shows the regression of a same-race dummy on peers' test scores. Column 1 shows that there is some correlation between peers' average test score and being assigned to a teacher of the same race in mathematics. Lower-quality peers are, as expected, more likely to be encountered when students are taught by a teacher of the same race. It is interesting that this effect disappears in column 2, where I control for a child fixed effect in a conditional logit regression. That is, looking at a given child moving from a teacher of the same race to a teacher of another race, peers' quality does not decline. Column 3 shows that controlling for teacher unobservables is not sufficient to control for peers' characteristics. Columns 4 to 6 present similar results for English teachers.

Table XVIII is another piece of evidence that suggests relative ranking is not the whole story. This table shows the results of the baseline regression of Table IV with an additional control for peers' average test score. These two tables are similar, and the hypothesis that the coefficients of interest (column 8) are equal between those two tables cannot be rejected at 95%. After controlling for peers' average test scores, child effects, teacher effects, and the test score, the result is that being assessed by a teacher of the same race increases test scores by 7.2% of a standard deviation in mathematics and by 4.4% of a standard deviation in English.

6 Conclusion

This paper uses a unique U.S. longitudinal data set that contains both teacher assessments and test scores. I assess whether teachers give better assessments to children of their race or gender. Controlling for child and teacher unobservables, I found that teachers give better assessments to children of their own race but not of their own gender. This effect might be due to white teachers giving lower grades to black and to hispanic children. It should be noted that a conservative interpretation of the results cannot determine whether teachers overassess or underassess pupils of their own race. Finally, results show that behavior is not significantly affected by same-race teachers conditional on test scores, teacher effects and student effects. This suggests that stereotype threat does not explain our main results.

This is the first large-scale analysis of teacher assessments versus test scores that uses U.S. elementary education data. Results highlight that teachers' races determine their perceptions of students' skills.

Controlled experiments on teachers' perceptions in U.S. classrooms are needed to assess both (i) how these perceptions affect children performance and (ii) how public policies (e.g. training) can change teachers' perceptions of their students.

References

Abowd, J., Creecy, R. and Kramarz, F. (2002), Computing person and firm effects using linked longitudinal employer-employee dataset.

Abowd, J. M., Kramarz, F. and Margolis, D. N. (1999), 'High wage workers and high wage firms', *Econometrica* **67**(2), 251–334.

Acemoglu, D. and Angrist, J. (2000), *How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws, NBER/Macroeconomics Annual*, NBER.

Aronson, J., Fried, C. B. and Good, C. (2002), 'Reducing the effects of stereotype threat on african american college students by shaping theories of intelligence', *Journal of Experimental Social Psychology* .

Card, D. and Sullivan, D. (1988), 'Measuring the effect of subsidized training programs on movements in and out of employment', *Econometrica* **56**(3), 497–530.

Clifford, M. and Walster, E. (1973), 'The effect of physical attractiveness on teacher expectations', *Sociology of Education* **46**(2), 248–258.

Clotfelter, C. T., Ladd, H. F. and Vigdor, J. L. (2005), 'Who teaches whom? race and the distribution of novice teachers', *Economics of Education Review* **24**, 377–392.

Coate, S. and Loury, G. C. (1993), 'Will affirmative-action policies eliminate negative stereotypes?', *American Economic Review* **83**(5), 1220–40.

Dee, T. S. (2004), 'Teachers, race, and student achievement in a randomized experiment', *The Review of Economics and Statistics* **86**(1), 195–210.

Dee, T. S. (2005*a*), 'A teacher like me: Does race, ethnicity, or gender matter?', *American Economic Review* **95**(2), 158–165.

Dee, T. S. (2005*b*), Teachers and the gender gaps in student achievement, NBER Working Papers 11660, National Bureau of Economic Research, Inc.

Efron, B. and Tibshirani, R. (1994), *An Introduction to the Bootstrap*, Chapman & Hall.

Fryer, R. G. and Levitt, S. D. (2006*a*), 'The black-white test score gap through third grade', *American Law and Economics Review* **8**(2).

Fryer, R. G. and Levitt, S. D. (2006*b*), Testing for racial differences in the mental ability of young children, NBER Working Papers 12066, National Bureau of Economic Research, Inc.

Heckman, J. J., Stixrud, J. and Urzua, S. (2006), 'The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior', *Journal of Labor Economics* **24**(3), 411–482.

Herrnstein, R. J. and Murray, C. (1994), *The Bell Curve: Intelligence and Class Structure in American Life*, Free Press.

Hughes, H. (1988), 'Psychological and behavioral correlates of family violence in child witnesses and victims', *American Journal of Orthopsychiatry* **58**, 77–90.

Kramarz, F., Machin, S. and Ouazad, A. (2007), What makes a test score? the respective contributions of pupils, peers and schools in achievement. mimeo.

Lavy, V. (2004), Do gender stereotypes reduce girls' human capital outcomes? evidence from a natural experiment, NBER Working Papers 10678, National Bureau of Economic Research, Inc.

Pelcovitz, D., Kaplan, S., Goldenberg, B., Mandel, F., Lehane, J. and Guarrera, J. (1994), 'Post-traumatic stress disorder in physically abused adolescents', *Journal of American Academy of Child and Adolescent Psychiatry* **33**, 305–312.

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

Rosenthal, R. and Jacobson, L. (1968), *Pygmalion in the Classroom*, Holt, Rinehart and Winston, New York.

Rothstein, J. (2008), Teacher quality in educational production: Tracking, decay, and student achievement. unpublished manuscript.

Rudner, L. M. and Schafer, W. D. (2001), 'Reliability', *ERIC Digest* .

Steele, C. and Aronson, J. (1998), 'Stereotype threat and intellectual test performance of african americans', *Journal of Personality and Social Psychology* **69**(5), 797–811.

Wheeler, S. and Petty, R. (2001), 'The effects of stereotype activation on behavior: A review of possible mechanisms', *Psychological Bulletin* .

	Mean		S.D.
Children's characteristics			
Male	0.503		(0.500)
White, non-hispanic	0.587		(0.492)
Black, African American	0.137		(0.344)
Hispanic, any race	0.157		(0.364)
Asian	0.057		(0.232)
Native Hawaiian, other Pacific Islander	0.012		(0.109)
American Indian or Alaska Native	0.018		(0.133)
More than one race	0.024		(0.154)
Test scores	50.296		(9.810)
Assessments	50.310		(9.877)
Teachers' characteristics			
Male	0.044		(0.205)
Race		— *See next table* —	
Age	42.255		(10.880)
Tenure	11.076		(9.273)
Experience at the grade level	8.536		(7.669)
Matching statistics			
Same-gender teacher	0.477		(0.499)
Same-race teacher	0.618		(0.486)
Sampled children per teacher	8.198		(5.914)
Same standards for everyone			
All fifth-grade teachers	0.12		(0.392)
White, non-hispanic	0.11		(0.501)
Black, African-American	0.19		(0.476)
American Indian or Alaska Native	0.07		(0.317)
Hispanic, any race	0.19		(0.475)
Native Hawaiian, other Pacific Islander	0.17		(0.353)
Asian	0.20		(0.397)
Male	0.15		(0.353)
Female	0.12		(0.324)

Some children and some teachers have a missing race variable; this case is treated as separate category and does not enter into the Same Race variable.

The bottom part of the table lists answers to the following question: "Which of the following best describes your evaluation and grading practices?". There are three possible choice, Same standards expect for special needs, Standards based on what they are capable of, and Same standards for everyone.

Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table I: Descriptive Statistics

	— All Teachers —				- English -	- Mathematics -
	Fall Kindergarten	Spring Kindergarten	Spring Grade 1	Spring Grade 3	Spring Grade 5	Spring Grade 5
Male	0.022	0.020	0.018	0.045	0.151	0.174
White, non-hispanic	0.739	0.740	0.607	0.561	0.742	0.737
Black, African American	0.062	0.064	0.054	0.052	0.080	0.086
Hispanic, any race	0.086	0.083	0.062	0.046	0.068	0.067
Asian	0.026	0.024	0.022	0.016	0.022	0.023
American Indian or Alaska Native	0.008	0.009	0.009	0.008	0.016	0.016
Native Hawaiian, other Pacific Islander	0.004	0.004	0.002	0.003	0.008	0.006
Number of teachers	3,132	3,388	5,046	6,093	4,735	4,697

Note: Some teachers have not reported their race; this case is treated as separate category and does not enter into the Same Race variable.

Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table II: Racial and Gender Diversity among Teachers from Kindergarten to Grade 5

	Mean	Teacher's Race: Same Race	Teacher's Race: Different Race	*Difference*
	(1)	(2)	(3)	*(4)=(2)−(3)*
Mathematics test scores				
Teacher assessments − Test scores				
Black, African American child	2.109	2.665	1.891	*0.774***
	(8.924)	(9.173)	(8.816)	*[0.248]*
White child	-1.040	-1.032	-1.122	*0.090*
	(8.921)	(8.849)	(9.655)	*[0.191]*
Hispanic, any race child	1.660	3.668	1.035	*2.634***
	(9.269)	(9.658)	(9.055)	*[0.240]*
English test scores				
Teacher assessments - Test scores				
Black, African American child	1.421	2.543	0.961	*1.583***
	(8.145)	(8.344)	(8.018)	*[0.184]*
White child	-0.585	-0.608	-0.323	*-0.286**
	(7.987)	(7.948)	(8.397)	*[0.145]*
Hispanic, any race child	1.326	4.220	0.687	*3.533***
	(8.693)	(9.087)	(8.472)	*[0.223]*

	Mean	Teacher's Gender: Same Gender	Teacher's Gender: Different Gender	*Difference*
	(1)	(2)	(3)	*(4)=(2)−(3)*
Mathematics test scores				
Teacher assessments - Test scores				
Male child	-0.678	-0.178	-0.701	*0.523*
	(9.224)	(8.770)	(9.244)	*[0.274]*
Female child	0.621	0.569	1.186	*-0.617***
	(8.921)	(8.941)	(8.685)	*[0.204]*
English test scores				
Teacher assessments - Test scores				
Male child	-0.309	0.159	-0.330	*0.489**
	(8.242)	(8.598)	(8.225)	*[0.229]*
Female child	0.410	0.333	1.255	*-0.922***
	(8.149)	(8.121)	(8.411)	*[0.165]*

Note: Standard deviations in parentheses (columns 1, 2, and 3), standard errors in brackets (column 4). The significance levels for standard errors are computed following a two-sample t-test with equal variances.
**, Significant at 1%; *, Significant at 5%.
Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table III: Descriptive Statistics on the Difference between Teacher Assessments and Test Scores

	Mathematics Teacher Assesments				English Teacher Assesments			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	OLS	Child f.e.	Teacher f.e.	Two-way f.e.	OLS	Child f.e.	Teacher f.e.	Two-way f.e.
Same-race Teacher	0.281*	0.704**	0.694**	0.711**	0.428**	0.413**	0.702**	0.435**
	(0.118)	(0.162)	(0.119)	(0.190)	(0.093)	(0.113)	(0.094)	(0.114)
Test score	0.591**	0.263**	0.588**	0.241**	0.659**	0.316**	0.669**	0.313**
	(0.004)	(0.009)	(0.004)	(0.009)	(0.003)	(0.006)	(0.003)	(0.005)
F-statistic	1,218.517	82.630	1,668.009	4.152	2,501.106	285.903	3,462.876	5.603
R-squared	0.348	0.666	0.540	0.786	0.436	0.699	0.553	0.773
Same-gender teacher	0.132	0.278	-0.083	-0.019	-0.221	-0.158	-0.215	-0.174
	(0.151)	(0.186)	(0.152)	(0.238)	(0.121)	(0.135)	(0.122)	(0.188)
Test score	0.591**	0.262**	0.587**	0.241**	0.659**	0.316**	0.668**	0.314**
	(0.004)	(0.009)	(0.004)	(0.005)	(0.003)	(0.006)	(0.003)	(0.006)
F-statistic	1,218.158	81.197	1,664.441	4.149	2,499.578	284.820	3,456.497	5.601
R-squared	0.347	0.665	0.539	0.786	0.436	0.699	0.552	0.773
Child controls	Yes	No	Yes	No	Yes	No	Yes	No
Teacher controls	Yes	Yes	No	No	Yes	Yes	No	No
Other controls	— Time dummies —							
Number of observations	48,065				67,855			

Note: 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.Standard errors are computed using bootstrapping in columns 4 and 8. Regressions are weighted using sampling design weights.
**, Significant at 1%; *, Significant at 5%; f.e.= fixed effects.
Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table IV: Do Same-Race or Same-Gender Teachers Give Better Assessments Conditional on Test Scores ?

	— Mathematics Teacher Assessment —		— English Teacher Assessment —	
	Fall Kindergarten	Spring Grade 5	Fall Kindergarten	Spring Grade 5
Boy	-1.131**	0.006	-1.796**	-3.208**
	(0.160)	(0.276)	(0.145)	(0.191)
Black, African American	-4.761**	-4.766**	-4.107**	-4.005**
	(0.227)	(0.458)	(0.205)	(0.312)
Hispanic, any race	-5.647**	-2.186**	-6.248**	-2.424**
	(0.210)	(0.359)	(0.192)	(0.251)
Asian	-1.885**	2.356**	-3.106**	1.613**
	(0.454)	(0.556)	(0.407)	(0.383)
Pacific Islander	-4.695**	-0.703	-5.078**	-2.803**
	(1.074)	(1.294)	(0.930)	(0.946)
Indian	-5.824**	-6.055**	-5.906**	-5.005**
	(0.604)	(0.990)	(0.543)	(0.717)
Observations	14,462	5,261	17,688	10,720
R-squared	0.07	0.04	0.08	0.05
F-statistic	117.41	26.19	161.11	68.39

Note: 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.
**, Significant at 1%; *, Significant at 5%.
Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table V: Gaps in Teacher Assessments from Kindergarten to Grade 5

	Child Fixed Effect		Teacher Fixed Effect	
	(1)	(2)	(3)	(4)
Male	-1.856**	-0.717**	1.221**	1.276**
	(0.094)	(0.083)	(0.369)	(0.378)
Black, African American	-1.669**	-0.594**	0.515	0.503
	(0.136)	(0.118)	(0.265)	(0.268)
Hispanic, any race	-1.382**	-0.604**	0.492	0.451
	(0.131)	(0.113)	(0.270)	(0.273)
Asian	-0.065	-0.403*	0.145	0.121
	(0.200)	(0.172)	(0.384)	(0.389)
Pacific Islander	-2.411**	-1.505**	-0.223	-0.340
	(0.461)	(0.395)	(1.036)	(1.039)
Indian	-2.914**	-1.901**	0.129	0.097
	(0.355)	(0.304)	(0.654)	(0.655)
Teacher's tenure				-0.044**
				(0.010)
Teacher's experience				0.021
				(0.012)
Child's behavior controls	No	Yes	-	-
Other controls	-	-	– Grade dummies –	
F-statistic	74.55	435.46	2.19	3.48
R-squared	0.03	0.29	0.00	0.01
Number of observations	20,131	20,131	5,496	5,268

Note: Male children's fixed effects are 18.6% of a standard deviation lower than female children's fixed effects when not controlling for the child's behavior. Male teachers' fixed effects are 1.2% of a standard deviation higher than female teachers' fixed effects.

**, Significant at 1%; *, Significant at 5%.

Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table VI: Analysis of Pupil and Teacher Effects

	Dependent Variable: Mathematics Teacher Assesments			Dependent Variable: English Teacher Assesments		
	Race of the Child			Race of the Child		
Race of the teacher	White, non-hispanic	Black, African American	Hispanic, any race	White, non-hispanic	Black, African American	Hispanic, any race
White, non-hispanic	Ref.	-0.616	-1.728**	Ref.	-1.110**	-1.480**
		(0.512)	(0.627)		(0.300)	(0.221)
Black, African American	-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		4.158			5.609	
R-squared		0.787			0.774	
Child fixed effects		Yes			Yes	
Teacher fixed effects		Yes			Yes	
Other controls		— Time dummies —			— Time dummies —	
Number of observations		48,065			67,855	

Note: 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.

**, Significant at 1%; *, Significant at 5%.

Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table VII: Do Same-Race Teachers Give Better Assessments? Analysis by Race

	Mathematics Teacher Assesments (1)	English Teacher Assesments (2)
	Two-way f.e.	Two-way f.e.
Male teacher - Female student	0.839	0.645*
	(0.504)	(0.261)
Female teacher - Male student	-0.339	-0.198
	(0.302)	(0.168)
Test score	0.241**	0.314**
	(0.013)	(0.009)
F-statistic	4.150	5.602
R-squared	0.786	0.773
Child Effects	Yes	Yes
Teacher Effects	Yes	Yes
Other controls	— Time dummies —	
Number of Observations	48,065	67,855

Note: 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.
**, Significant at 1%; *, Significant at 5%; f.e.= fixed effects.
Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table VIII: Do Same-Gender Teachers Give Better Assessments? Analysis by Gender

	Same Race Teacher			
Boy	-0.003	-0.003		
	(0.004)	(0.004)		
Black African	-0.638**	-0.638**		
	(0.005)	(0.005)		
Hispanic, Any Race	-0.683**	-0.683**		
	(0.005)	(0.005)		
Asian	-0.828**	-0.828**		
	(0.008)	(0.008)		
Native Hawaiian, other Pacific Islander	-0.814**	-0.816**		
	(0.018)	(0.018)		
American Indian or Alaska Native	-0.740**	-0.740**		
	(0.013)	(0.013)		
Approaches to learning		0.001	0.001	-0.000
		(0.001)	(0.001)	(0.001)
Self-control		0.001	-0.000	-0.000
		(0.001)	(0.001)	(0.001)
Interpersonal skills		-0.002**	-0.001	-0.001
		(0.001)	(0.001)	(0.001)
Externalizing Problems Behavior		-0.000	0.000	0.000
		(0.000)	(0.001)	(0.001)
Internalizing Problems Behavior		-0.000	0.000	0.000
		(0.000)	(0.000)	(0.001)
Child Fixed Effect	No	No	Yes	Yes
Teacher Fixed Effect	No	No	No	Yes
F Statistic	3,397.235	1,854.790	2.049	5.868
R Squared	0.528	0.528	0.809	0.903
Number of Observations	36,465	36,465	36,465	36,465

Note: Behavioral measures are reported by the teacher (Teacher Social Rating Scale). Standard errors are bootstrapped in column 4. Observations are for the matching of children to mathematics teachers. Results are similar for English teachers. **, Significant at 1%; *, Significant at 5%.
Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table IX: The Matching of Teachers to Children: Race, Gender and Behavior

	Mathematics Teacher Assessments								
	$\theta = 0.0$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.4$	$\theta = 0.5$	$\theta = 0.6$	$\theta = 0.7$	$\theta = 0.8$
Same-race teacher	0.711**	0.620*	0.506**	0.360	0.164	-0.111	-0.523**	-1.213**	-2.597**
	(0.177)	(0.253)	(0.132)	(0.259)	(0.196)	(0.243)	(0.160)	(0.266)	(0.287)
Corrected test score	0.241**	0.268**	0.301**	0.345**	0.402**	0.483**	0.605**	0.809**	1.216**
	(0.006)	(0.011)	(0.009)	(0.011)	(0.014)	(0.020)	(0.028)	(0.030)	(0.054)
	English Teacher Assessments								
	$\theta = 0.0$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.4$	$\theta = 0.5$	$\theta = 0.6$	$\theta = 0.7$	$\theta = 0.8$
Same-race teacher	0.435**	0.327	0.193	0.021	-0.208	-0.525**	-0.997**	-1.763**	-3.197**
	(0.099)	(0.170)	(0.120)	(0.102)	(0.131)	(0.130)	(0.130)	(0.154)	(0.113)
Corrected test score	0.313**	0.348**	0.391**	0.446**	0.520**	0.622**	0.772**	1.016**	1.470**
	(0.009)	(0.008)	(0.007)	(0.011)	(0.007)	(0.011)	(0.018)	(0.022)	(0.034)

Note: 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. The corrected test score is such that $\tilde{y}_{i,f,t} = \theta \cdot E[y_{.,f,t} \mid \text{Same Race}] + (1-\theta) \cdot y_{i,f,t}$, where i indexes children, f=field (maths or English), and t goes from fall kindergarten to spring grade 5.

**, Significant at 1%; *, Significant at 5%.

Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table X: What Amount of Measurement Error in Test Scores Could Explain the Results?

	Race				Gender	
	(1)	(2)	(3)	(4)	(5)	(6)
Correlation across fields	All children	Black	Non-hispanic white	Hispanic	Male	Female
Test scores	0.739	0.745	0.709	0.722	0.740	0.753
Teacher assessments	0.803	0.812	0.794	0.795	0.803	0.809
Δ = TS − TA	0.532	0.530	0.528	0.526	0.512	0.552
Number of observations	45,923					

TA = Teacher assessment; TS = Test score.

Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table XI: Correlation of Teacher Assessment—Test Score Gap between Topics

	Dependent Variable			
	(1)	(2)	(3)	(4)
	Learning	Control	Externalizing Problems	Internalizing Problems
Same-race teacher	-0.052	0.276	-0.121	-0.004
	(0.245)	(0.230)	(0.246)	(0.301)
Test score	0.119**	0.054**	-0.053**	-0.042**
	(0.015)	(0.015)	(0.013)	(0.014)
F-statistic	3.925	2.845	3.952	2.230
R-squared	0.777	0.716	0.778	0.664
Child fixed effects	Yes	Yes	Yes	Yes
Teacher fixed effects	Yes	Yes	Yes	Yes
Other controls	— Time dummies —			
Number of Observations	48,037			

Note: 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.
**, Significant at 1%; *, Significant at 5%.
Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table XII: Behavior and Same-Race Teaching

Mobility Pattern	Count		White	Black, African American	Hispanic
00000	6759	(31.6 %)	0.04	0.25	0.35
00001	345	(1.6 %)	0.11	0.39	0.28
00010	283	(1.3 %)	0.14	0.39	0.35
00011	143	(0.7 %)	0.47	0.29	0.17
00100	580	(2.7 %)	0.25	0.33	0.27
00101	172	(0.8 %)	0.40	0.32	0.22
00110	169	(0.8 %)	0.52	0.18	0.28
00111	307	(1.4 %)	0.82	0.09	0.08
01000	260	(1.2 %)	0.68	0.19	0.10
01001	18	(0.1 %)	0.78	0.11	0.11
01010	30	(0.1 %)	0.83	0.13	0.00
01011	45	(0.2 %)	0.98	0.00	0.02
01100	155	(0.7 %)	0.86	0.08	0.05
01101	45	(0.2 %)	0.96	0.04	0.00
01110	116	(0.5 %)	0.91	0.05	0.04
01111	360	(1.7 %)	0.96	0.01	0.03
10000	653	(3.1 %)	0.75	0.11	0.11
10001	20	(0.1 %)	0.65	0.10	0.20
10010	29	(0.1 %)	0.66	0.14	0.10
10011	27	(0.1 %)	0.89	0.00	0.04
10100	26	(0.1 %)	0.73	0.00	0.23
10101	7	(0.0 %)	0.86	0.00	0.14
10110	18	(0.1 %)	0.72	0.06	0.22
10111	19	(0.1 %)	0.84	0.05	0.05
11000	2489	(11.6 %)	0.70	0.14	0.12
11001	236	(1.1 %)	0.63	0.15	0.18
11010	342	(1.6 %)	0.74	0.13	0.11
11011	473	(2.2 %)	0.91	0.03	0.05
11100	1626	(7.6 %)	0.81	0.06	0.10
11101	705	(3.3 %)	0.85	0.05	0.08
11110	1188	(5.5 %)	0.90	0.04	0.06
11111	3764	(17.6 %)	0.97	0.02	0.01

Note: '00101' means that the child had a teacher of the same race in spring first grade and spring fifth grade, and a teacher of a different race in fall and spring kindergarten and spring third grade.
Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table XIII: Mobility Patterns in Mathematics: Same-Race Teacher vs. Other-Race Teacher

	Progress in Mathematics Teacher Assesments (1)	Progress in English Teacher Assesments (2)
	Two-way f.e.	Two-way f.e.
Moving to same-race teacher	0.803	0.451
	(0.603)	(0.321)
Progress in test score	0.029	0.176**
	(0.029)	(0.009)
F-statistic	1.250	1.128
R-squared	0.830	0.510
Child effects	Yes	Yes
Teacher effects	Yes	Yes
Other controls	— Time dummies —	
Number of observations	22,073	44,471

Note: 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.
**, Significant at 1%; *, Significant at 5%; f.e.= fixed effects.
Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table XIV: Do Same-Race or Same-Gender Teachers Give Better Assessments? — Robustness Check

	Mathematics Test Scores (1)	English Test Scores (2)
	Two-way f.e.	Two-way f.e.
Same-race teacher	-0.151	-0.060
	(0.140)	(0.103)
Teacher assessment	0.118**	0.212**
	(0.005)	(0.004)
F-statistic	9.540	8.865
R-squared	0.894	0.843
Same-gender teacher	0.537**	0.346*
	(0.192)	(0.148)
Teacher assessment	0.118**	0.212**
	(0.005)	(0.002)
F-statistic	9.546	8.867
R-squared	0.894	0.843
Child Effects	Yes	Yes
Teacher Effects	Yes	Yes
Other controls	— Time dummies —	
Number of observations	48,065	67,855

Note: 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.
**, Significant at 1%; *, Significant at 5%; f.e.= fixed effects.
Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table XV: Do Same-Race or Same-Gender Teachers Give Better Assessments? — Falsification check

	Rank in Mathematics Teacher Assesments				Rank in English Teacher Assesments			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	OLS	Child f.e.	Teacher f.e.	Two-way f.e.	OLS	Child f.e.	Teacher f.e.	Two-way f.e.
Same-race teacher	0.097**	0.058	0.020	-0.026	0.119**	0.060	0.097**	0.018
	(0.031)	(0.051)	(0.030)	(0.045)	(0.024)	(0.034)	(0.023)	(0.021)
Rank in test scores	0.791**	0.632**	0.700**	0.458**	0.826**	0.652**	0.767**	0.525**
	(0.006)	(0.011)	(0.006)	(0.008)	(0.004)	(0.008)	(0.004)	(0.009)
F-statistic	1,537.421	790.854	1,463.337	6.127	2,499.253	1,217.505	2,742.233	9.233
R-squared	0.636	0.811	0.678	0.845	0.704	0.829	0.723	0.849
Same-gender teacher	0.052	-0.036	0.015	-0.029	0.024	0.027	0.031	0.051
	(0.038)	(0.062)	(0.043)	(0.059)	(0.028)	(0.041)	(0.032)	(0.061)
Rank in test scores	0.791**	0.633**	0.700**	0.458**	0.827**	0.652**	0.767**	0.525**
	(0.006)	(0.011)	(0.006)	(0.011)	(0.004)	(0.008)	(0.004)	(0.006)
F-statistic	1,537.080	790.681	1,462.631	6.127	2,496.582	1,216.795	2,741.014	9.234
R-squared	0.636	0.811	0.678	0.845	0.704	0.829	0.722	0.849
Child controls	Yes	No	Yes	No	Yes	No	Yes	No
Teacher controls	Yes	Yes	No	No	Yes	Yes	No	No
Other controls	— Time dummies —							
Number of observations	48,065				67,855			

Note: 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.**, Significant at 1%; *, Significant at 5%.Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table XVI: Grading on a Curve: Do Same-Race or Same-Gender Teachers Give Better Assessments?

	Same-Race Teacher: Mathematics			Same Race Teacher: English		
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	Child f.e.	Teacher f.e.	OLS	Child f.e.	Teacher f.e.
Peers' average test score						
Point estimate	-0.016**	-0.002	-0.009**	-0.011**	0.003	-0.009**
S.E.	(0.001)	(0.005)	(0.001)	(0.001)	(0.004)	(0.001)
Odds ratio	0.984**	0.998	0.991**	0.989**	1.003	0.991**
Test score						
Point estimate	-0.004**	-0.002	-0.012**	-0.004**	-0.002	-0.012**
S.E.	(0.001)	(0.005)	(0.002)	(0.001)	(0.004)	(0.002)
Odds ratio	0.996**	0.998	0.988**	0.996**	0.998	0.989**
Chi squared	25757.001	403.529	22933.218	38159.316	558.057	36282.184
Pseudo R-squared	0.422	0.067	0.591	0.445	0.052	0.592
Child controls	Yes	No	Yes	Yes	No	Yes
Teacher controls	Yes	Yes	No	Yes	Yes	No
Other controls	— Time dummies —					
Number of observations	46,597			65,542		

Note: 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.**, Significant at 1%; *, Significant at 5%.Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table XVII: Is Peer Quality Correlated with Same-Race Teaching?

	Mathematics Teacher Assesments (1)	English Teacher Assesments (2)
	Two-way f.e.	Two-way f.e.
Same-race teacher	0.718**	0.438*
	(0.255)	(0.189)
Test score	0.240**	0.316**
	(0.009)	(0.006)
F-statistic	4.155	5.615
R-squared	0.787	0.773
Peers controls	Yes	Yes
Child controls	Yes	Yes
Teacher controls	Yes	Yes
Other controls	— Time dummies —	
Number of Observations	48,065	67,855

Note: 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.**, Significant at 1%; *, Significant at 5%.Source: Early Childhood Longitudinal Study, Kindergarten cohort of 1998–1999.

Table XVIII: Controlling for Peer Quality