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

Amine Ouazad[*] [†]

October 2007

Abstract

Racial and gender gaps widen from kindergarten to fifth grade and several authors point out that these gaps might be due to teacher quality. I look at one particular aspect of teachers' behavior. Do teachers give better grades to children of their own race, ethnicity or gender? A U.S. child-teacher matched longitudinal dataset 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 conditionally on test scores, child effects and teacher effects. The effect of same race or same gender teaching on assessments is therefore identified using pupil mobility. 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 race, but not significantly higher to children of their gender. Also, this effect comes from the differential assessments given to non-hispanic black and hispanic children. Non-hispanic black teachers give significantly higher assessments to non-hispanic black children, and hispanic teachers give significantly higher assessments 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: I2

[*]I am indebted to Roland Benabou, Marc Gurgand, Xavier d'Haultfoeuille, Caroline Hoxby, Francis Kramarz, Eric Maurin, Jesse Rothstein, 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 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 dataset. The author acknowledges financial support from the Economics of Education and Education Policy in Europe network, Princeton University and CREST-INSEE.

[†]Centre for Economic Performance, London School of Economics and Paris School of Economics. Address: Centre for Economic Performance, Houghton Street, London WC2A 2AE, United Kingdom. Tel: +44(0)20 7955 7800. Email: a.ouazad@lse.ac.uk Website for econometric programs: http://www.pse.ens.fr/junior/ouazad/

1 Introduction

Educational policies at the federal and the state level increasingly rely on test scores: test scores are indeed comparable across classes, schools and, potentially, states. As standardized testing becomes widespread, polls reveal growing discontent among parents. In 1997, 24% of public school parents felt there was too much emphasis on achievement testing (Gallup polls quoted by Rose & Gallup (2006)). In 2000, this fraction jumped to 34% , and it went on rising up to 45% in 2006. And the debate on the reauthorization of No Child Left Behind has sparked criticism of the current testing regime. A letter addressed to Congress signed by practitioners and scholars in education asserts that "Performance based assessments, often locally controlled and involving multiple measures of achievement, offer a way to move beyond the limits and negative effects of standardized examinations" (Wood, Darling-Hammond, Neil & Roschewski 2007). Relying on subjective assessments of children's skills is indeed the alternative to standardized testing, since most teachers fill school records that include comments on the child's ability or behavior. And important decisions such as tracking, special education and ability are partly based on subjective assessments. Teachers' priors, beliefs and behavior may be based on what other teachers reported.

Table 1 shows how 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%, 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.

Do teachers hold all children to the same standards, regardless of race and gender? I use a unique US longitudinal dataset that combines test scores and teacher assessments of children' skills in elementary education. I can therefore compare the difference between test scores and teacher assessments when the same child experiences same race teachers and when he has a teacher of a different race. I can also look at this difference for the same teacher when assessing same race children and children of different races. Combining these two identification strategies, I estimate the effect of same race and same gender teaching on assessments, conditionally on test scores, child and teacher fixed effects. This addresses three potential identification issues: firstly, children of different genders and races may behave differently in the classroom and during examinations, e.g. differential effect of testing on boys and girls, stereotype threat effects (Steele & Aronsson 1998); secondly, teacher assessments may capture skills that are not captured by test scores; finally, some teachers may give higher average assessments, and this can be correlated with

	Which of the following best describes your evaluation and grading practices?		
	Same standards except for special needs	Standards based on what they are capable of	Same Standards for everyone
All Teachers	0.70	0.17	0.12
White, Non Hispanic	0.71	0.17	0.11
Black, African American	0.59	0.22	0.19
American Indian or Alaska Native	0.71	0.22	0.07
Hispanic, Any Race	0.70	0.11	0.19
Native Hawaiian, other Pacific Islander	0.52	0.31	0.17
Asian	0.66	0.15	0.20
Male	0.68	0.17	0.15
Female	0.71	0.17	0.12

Source: Early Childhood Longitudinal Study, Kindergarten Cohort of 1998/1999.

Table 1: Fifth Grade Teachers' Self-Reported Grading Practices

child characteristics.

The dataset is the Early Childhood Longitudinal Study, Kindergarten cohort of 1998-1999, collected by the National Center for Education Statistics at the US Department of Education. It is the first large scale US study that follows a cohort of children from kindergarten entry to middle school. This is therefore the first paper that looks at the discrepancy between test scores and teachers' perception of child skills using a representative longitudinal sample of US children in elementary education. Important findings are that teachers tend to give better assessments to children of their race and ethnicity, but not significantly higher assessments to children of their gender. Moreover this result is mainly due to grades given to non-hispanic black and hispanic (any race) children. Non-hispanic black teachers give better assessments to non-hispanic black children and hispanic teachers give better assessments 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. Moreover, measurement error checks show that only a large amount of measurement error can explain results. The estimates are also robust to falsification checks in which test scores are regressed on test scores rather than test scores on teacher assessments. Finally, I show that even if relative ranking and racial *de facto* segregation could be a potential explanation, controlling for peers' test scores does not change the results.

The analysis of this paper is comparable to Lavy (2004). Lavy's paper uses high school matriculation exams in Israel. Comparison of blind versus non blind test scores showed that boys are likely to be overassessed in all subjects. Moreover, the size of the bias was very sensitive to teachers' characteristics suggesting that teachers' behavior is causing grade discrimination. This paper differs from Lavy (2004) in two ways. First, I compare subjective assessments and test scores, where subjective assessments are based on classroom behavior and coursework. Second, I use the longitudinal dimension of the dataset to study teacher-student interaction. My paper's identification strategy controls for both child and teacher fixed effects. In Lavy (2004), if tough teachers are more likely to grade boys, the effect of non-blind assessments on boys' test scores could be overestimated. I control for this effect in the ECLS-K.

This paper is also related to a small scale experiment on fifth grade teachers in the state of Missouri. Clifford & Walster (1973) sent report cards to teachers. Theses cards included child records randomly matched to photographs, and teachers were asked to assess child ability. They found a significant effect of physical attractiveness on assessments, but no effect of gender. This study nevertheless raises a number of issues. It is not clear whether this result on Missouri fifth grade teachers may be relevant to assess discrimination in a representative U.S. classroom: teachers assess students they do not know on the basis of randomly generated school records. This paper's analysis on the ECLS-K provides a large scale analysis of teachers' assessments in U.S. elementary education.

Better teacher assessments may have different effects. 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, in a similar fashion as in Coate & 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. On the contrary, assessing grading discrimination requires a dataset such as the ECLS-K. Some evidence of the effect of same race or same gender teaching is available though: 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 shows that same gender teaching increases the test scores of 8th grade children (Dee 2005*b*).

The psychological and educational literature has debated on the issue of the effect of teacher expectations at least since the Pygmalion experiment (Rosenthal & 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. In this paper, I show that teacher have better perceptions of their students' skills when the teacher and his student share the same race and ethnicity.

This evidence is in line with previous results on racial gaps: even though there is no significant difference in the cognitive performance of white and black children aged eight months to a year (Fryer & Levitt 2006*b*), the gaps grow along the curriculum (Fryer & Levitt 2006*a*). Fryer & Levitt (2004) suggest that school quality may be part of the explanation. And since teacher quality is likely to be the most important determinant of school quality (Hanushek & Kain 2005), teachers might be part of the story for the growing racial gaps.

The rest of the paper is structured as follows. Section 2 presents the Early Childhood Longitudinal Study. It provides a description of racial and gender diversity in U.S. elementary education as well as descriptive statistics on teacher assessments and test scores. Section 3 explains main identification issues, the identification strategy and baseline results. Section 4 checks the robustness of the results, and section **??** goes deeper by looking at the results race by race and gender by gender. Section 5 shows that assessment rankings are not affected by teacher-pupil racial interactions in the classroom, but 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 US 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 one and in the spring of grades three and five. The study's last followup will be eighth grade. Followups 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 dataset to strengthen the representativeness of the survey. In the spring of 1999, part of the schools that had previously declined participation were included. In the spring of grade one, new children were included; this made the cross sectional sample representative of grade one children. Children have then been followed in the spring of grade three and five.

This paper's empirical analysis uses the restricted use version of the ECLS-K which contains the race and the gender of both the teacher and his pupils. Some observations with missing data on basic variables (test scores, subjective assessments, teachers' and children' race and gender) were deleted; weights provided by the survey's designers correct for the subsampling of movers. Race and ethnicity questions for the teacher were combined to match the categories of the child's race question; therefore

'Hispanic, Any Race' is a separate category. Same race should be subsequently read as 'same race, same ethnicity'[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 between 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 fill one questionnaire per child. Teacher assesments of children's skills, also called the Academic Rating Scale, are separated into three areas: (i) Language and Literacy (ii) General Knowledge (ii) Mathematical Thinking. I will use English, ie Language and Literacy, and mathematics assessments, ie mathematical thinking. The instructions make it clear that it is not a test ans should not be administered directly to the child. In English and Maths, 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, that will be 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 $p_i(\theta) = c_i + (1 - c_i)/(1 + e^{-Da_i(\theta - b_i)})$, where a_i,b_i and c_i are question specific parameters and θ is child ability.

the learning environment. The Self-Control Scale indicates the childs 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 childs 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, and sadness.

Basic children's characteristics are depicted in table 2. 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 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, therefore the potential advantages of same race or same gender teaching will mostly be felt by female or white children. Tables 2 and 3 show that only 4.4% of teachers are male. 47.7% of children are matched with a teacher of the same gender. However, the fraction of male teachers increases over time. 2.2% of fall kindergarten teachers are male , but it jumps to 15.1% among grade five English teachers and 17.4% among grade five mathematics teachers.

Teachers are also mostly white, with table 3 revealing that 73.9% of teachers are non hispanic white in fall kindergarten. This fraction decreases along the curriculum but goes up in grade 5. Most of minority teachers are either hispanic (of any race) or black, African Americans. They predominantly teach to minority children, with minority teachers' classrooms made up on average of 81.4% minority children. Column (1) of table 14 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 between 65 and 67% less likely to be taught by a teacher of the same race. It goes down to 83% for Asian children.

A first taste of the forthcoming results is shown in the descriptive statistics of tables 4 to 7. Let me start with mathematics. The difference between test scores and teacher assessments is higher when matched with a teacher of the same race for Black, African American children (7% of a standard deviation), Hispanic, Any Race children (26.3% of a standard deviation), and American Indian or Alaska Native children (42.6% 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 Black, African American children, 35.3% for Hispanic, Any Race children, 33.2% of a standard deviation for American Indian or Alaska Native children. The difference is slightly negative for white children in English, but this effect will disappear when controlling for confounding effects. Descriptive statistics for same gender vs different gender pairings do not display the same clear-cut figures. 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 seen as causal as they do not control for potentially confounding effects. I describe them in the next section.

3 Identification and Results

3.1 Baseline specification

Descriptive statistics suggest that in most minority groups, the teacher assessment-test score gap is higher when matched to a teacher of the same race (tables 4 and 5). This may not be interpreted as a causal effect for a number of reasons.

Firstly, teachers may capture skills that are not captured by test scores. The description of the dataset 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 possible answers, ie 'Not Yet', 'Beginning', 'In Progress', 'Intermediate', 'Proficient'.

Secondly, boys, 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.

Thirdly, 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 correlated with being of the same race or same gender as your children; 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 the same gender. Formally,

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

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

Where $a_{i,f,t}$ is the teacher assessment of child i in field f = English, Maths, in period t. t runs from fall kindergarten to spring grade 5. $y_{i,f,t}$ is the test score, $u_{i,f}$ the child effect of child i in field f. $\mu_{J(i,f,t)}$ is the teacher effect. Same Race$_{i,f,t}$ (Same Gender$_{i,f,t}$) takes value 1 when matched with a teacher of the same race (gender), 0 otherwise.

$u_{i,f}$ captures non time varying individual characteristics that may have an effect on assessments regardless of the teacher. This for instance may capture behavior, that teachers may on average include in their assessments. Boys may also 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, the OLS estimates of α_g and α_r might be biased. The teacher effect $\mu_{J(i,f,t)}$ therefore captures these permanent average differences between teachers[3].

The model is estimated using a preconditioned conjugate gradient method described in Abowd, Creecy & Kramarz (2002)[4]. All estimations have converged with a numerical precision of 10^{-15}. Standard errors were obtained by bootstrapping using simple random sampling.

As in Abowd, Kramarz & Margolis (1999) and Kramarz, Machin & Ouazad (2007), children moving from/to a same race teacher identify the effect of same race assessments and same gender teacher on assessments conditionally on test scores. The identification of specifications 1 and 2 therefore requires sufficient and exogenous mobility[5].

Exogenous mobility is best understood when comparing the progress of a child in terms of assessments to the progress of the child in terms of test scores. Let us therefore take the first difference of specifications

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

[4] I have developed a set of STATA packages available on the web at http://repository.ciser.cornell.edu/viewcvs-public/cg2/branches/stata/, or by typing net from http://repository.ciser.cornell.edu/viewcvs-public/cg2/branches/stata/ on the command line.

[5] Sufficient mobility can be properly defined. In the same way 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. All teachers need to be in the same connex component of the mobility graph. It is then possible to identify the relative toughness in grading of all teachers.

(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 assessments 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, conditionally 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 conditionally on the other covariates. Section **??** 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 significantly better assessments to pupils of their gender. The effect is sizeable: it is between 1/10 and 1/5 of the black-white teacher assessment gap, and around 1/3 of the hispanic-non-hispanic white teacher assessment gap.

Baseline results are presented in table 8. OLS estimates indicate that children who are assessed by same race teachers also have higher maths 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 above. Column (2) gives the estimate when controlling for child effects. The estimate is higher than the baseline OLS one, which suggests that the child fixed effect is negatively correlated with same race pairings. Most teachers are female nonhispanic white, thus either on average all teachers give lower assessments to white children or, white kids respond differently when in the classroom and 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. This may be due to the fact that teachers who give lower assessments are matched with children of the same race. Again a majority of teachers are female nonhispanic white, and a possible story is that these teachers are tougher than teachers of other races & ethnicities.

Finally, column (4) gives the estimate when controlling for both children and teacher fixed effects. The estimate is fairly similar to the estimates of columns (2) and (3). Column (4) is my preferred estimate for the effect of same race matching on assessments conditionally on test scores. It indeed addresses the three important identification issues described above. On average, children who are assessed by a teacher of the same race have a higher mathematics assessment, by 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 assessments, 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 maths assessments. Column (8) shows the estimate when controlling for both children and teacher fixed effects. Interestingly, 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, conditionally on test scores and children and teacher fixed effects.

The gender and racial gaps in teacher assessments are shown on table 9. This table is useful to compare the gaps in assessments to the magnitude of grade discrimination. In mathematics, the effect of same race assessments is around 7% of a standard deviation, and that is around 1/3 of the black-white teacher assessment gap and 1/5 of the non-hispanic white-hispanic teacher assessment gap. In English, the effect of same race assessments is around 4.1% of a standard deviation, which is around 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 and teacher fixed effects for the same race regressions in English and mathematics can be analyzed. Child fixed effects are interpreted as: (i) differential behavior in the classroom and during tests (ii) unobserved characteristics that teachers may on average include in their assessment (iii) average grading discrimination. Column (1) of table 10 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 10 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 10 shows 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; this even though the share of minority teachers has steadily declined in the last decades.

3.4 Do White Teachers Give Better Assessments to White Pupils?

Results have suggested that teachers give higher grades to children of their own race conditionally on test scores and children' and teachers' constant characteristics. What races drive the result? 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 race. This will allow for heterogeneous effects, race by race.

The specification is similar to baseline specification 1.

$$\begin{aligned} a_{i,f,t} &= \mu_{J(i,f,t)} + \delta y_{i,f,t} \\ &+\alpha_{\text{Black, African American}}(\text{Same Race - Black, African American})_{i,f,t} \\ &+\alpha_{\text{White}}(\text{Same Race - White, Non-Hispanic})_{i,f,t} \\ &+\ldots \\ &+u_i + \varepsilon_{i,f,t} \end{aligned} \tag{5}$$

Pupil i's assessment $a_{i,f,t}$ in field f = English, Maths in period t depends on test scores $y_{i,f,t}$, a set of $(\text{Same Race})_{i,f,t,\text{White, Non-Hispanic}}$, $(\text{Same Race})_{i,f,t,\text{Black, African American}}$, etc. dummies, child effects u_i and teacher effects $\mu_{J(i,f,t)}$.

Results are presented in table 18 for mathematics assessments and in table 19 for English assessments, and they look similar in both fields. A striking stylized fact is that the effect of same race teachers on assessments is mostly due to black and hispanic teachers. Being matched with a Black teacher increases Black children' assessments by 11% of a standard deviation in English, by 9% of a standard deviation in mathematics. For hispanic children, the effects are respectively 12% in English and 14.6% in mathematics.

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 conditionally on test scores and children' 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 male teacher–male child and female teacher–female child pairings.

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

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

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

Empirical results presented in table 17 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. These results suggest that teachers in minority groups – male teachers, black teachers, hispanic teachers – are more likely to favor children of their own race or gender.

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, for instance, (i) the behavior is the child is implicitly part of the teacher assessments and (ii) his behavior makes him 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 domestice 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 & Guarrera 1994). Family events may therefore drive behavioral changes. Moreover results from the analysis of a dataset of the North Carolina Department of Public Instruction suggest that novice teachers are assigned to classrooms in a way that disadvantages black students (Clotfelter, Ladd & Vigdor 2005).

Columns (3) and (4) of table 14 show that there is no significant effect of behavior on the probability of being matched with a teacher of the same race is not significant. There is little correlation between behavior and same race teaching when controlling for teacher fixed effects, and it disappears 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.

4.2 Minority children are more likely to be matched to a teacher of their own race

Minority pupils are less likely to be matched to a teacher of the same race. They are also less likely to move from a teacher of a different race to a teacher of the same 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 13 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, while 25% of Black, African American children have always been taught by a teacher of a different race. Column (1) of table 14 shows that while gender is not correlated with same race teaching, minority pupils are less likely to be matched with a teacher of the same race in the early years of elementary education. There is indeed a link between race and mobility patterns.

I introduce child and teacher fixed effects in the first differenced equation. That allows for some correlation between mobility and children's observed and unobserved characteristics.

$$\Delta a_{i,f,t} = \delta \Delta y_{i,f,t} + \alpha_r \Delta \text{Same Race}_{i,f,t} + u_i + \mu_{J(i,f,t)} + \nu_{i,f,t} \quad (7)$$

$$\Delta a_{i,f,t} = \delta \Delta y_{i,f,t} + \alpha_g \Delta \text{Same Gender}_{i,f,t} + u_i + \mu_{J(i,f,t)} + \nu_{i,f,t} \quad (8)$$

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

Table 15 show the results for the estimation of specifications 7 and 8. 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 maths assessments. The effect is not significant for mathematics; it is very similar to the baseline estimates (7% in column (4) of table 8, and 8% in column (4) of table 15). The estimate for English assessments is both significant and very close to the baseline estimate, with children paired with a teacher of the same race having 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.

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

Baseline results suggest that teachers give significantly higher assessments to children of their race. However other stories could explain the result. Teacher assessments may be driving test scores, such that expectations actually have an impact on educational outcomes; or teachers could try to predict the outcome of examinations. In both cases, the effect of same race teaching goes from teacher assessments to test scores, and not the reverse. The following specifications test for potential reverse causality, and empirical results suggest that these stories are not relevant.

In this falsification test, test scores and teacher assessments are therefore reversed, assessments explain test scores rather than the other way round.

$$y_{i,f,t} = \mu_{J(i,f,t)} + \delta a_{i,f,t} + u_{i,f} + \alpha_r \text{Same Race}_{i,f,t} + \varepsilon_{i,f,t} \quad (9)$$

$$y_{i,f,t} = \mu_{J(i,f,t)} + \delta a_{i,f,t} + u_{i,f} + \alpha_g \text{Same Gender}_{i,f,t} + \varepsilon_{i,f,t} \quad (10)$$

Notations are as in the baseline specifications 1 and 2. Results are presented in table 16: while the OLS estimates are significantly negative, the effect of same race teaching becomes non-significant when adding a child fixed effect, in both mathematics and English specifications. The effect is also non-significant when controlling for both child and teacher fixed effects. This suggests that it is unlikely that reverse causality is an alternative story.

4.4 Does Measurement Error Explain the Results?

Test scores of multiple choice questionnaires are usually noisy measures of underlying ability (Rudner & 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; items may be either not sufficiently difficult or too difficult for the child. Noise may be also be due to children' behavior, such as sleep patterns, illnesses, careless errors when filling the questionnaire, 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 nonhispanic white, and most minority teachers are either hispanic or black, African American. The 'same race' variable will therefore be correlated with the gap between white and black 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 8.

Let me explain that in formal terms. Consider that test scores are noisy measures of the child's underlying ability:

$$y_{i,t} = y^*_{i,t} + \nu_{i,t} \quad (11)$$

I assume that measurement error is classical, ie $\nu_{i,t}$ is not correlated with ability. In other words, this 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:

$$a_{i,f,t} = \mu + \delta y^*_{i,f,t} + \alpha_r \text{Same Race}_{i,f,t} + \varepsilon_{i,f,t} \tag{12}$$

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 estimates equation 12 by regressing on $y_{i,f,t}$. Then, both the estimate of δ and the estimate of α_r will be biased.

$$\hat{\alpha}_{r,OLS} = \alpha_r + \delta \cdot \lambda\theta \tag{13}$$

Where

$$\theta = Var(\nu)/[Var(\nu) + Var(y^*)] \tag{14}$$

$$\lambda = \frac{Cov(\text{Same Race}, y^*)}{Var(\text{Same Race})(1 - Corr(\text{Same Race}, y^*)^2)} \tag{15}$$

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 (Phelps 2007).

This shows that measurement error can lead to the overestimation of the effect of same race teachers on assessments conditionally on test scores. Given the knowledge of the relative size θ of the measurement error, one could estimate the unbiased effect of same race teachers on assessments. Indeed, build the following corrected value of the test score:

$$\tilde{y}_{i,f,t} = \theta \cdot E[y_{\cdot,f,t}|SameRace] + (1 - \theta) \cdot y_{i,f,t} \tag{16}$$

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 conditionally on test scores.

But the size of measurement error is unknown, therefore I will 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 11. 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. In English, a measurement error of about 20% is required to make the coefficient nonsignificant, whereas a measurement error of about 30 to 40% cancels the point estimate. In a word,at least 20 to 30% of measurement error would be necessary to explain the coefficient. Even though this result does not exclude a potential confounding effect of measurement error, it suggests that only a large amount of measurement error would alter our conclusions.

4.5 Are the Teacher Assessment–Test Score Gaps Correlated Across Topics?

The analysis has been carried out so far 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, they should be correlated across topics. From kindergarten to third grade, it is indeed the same teacher who fills both teacher assessment forms in English and mathematics. On the other hand, if the difference between teacher assessments and test scores is only measurement error, 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 12 shows the correlation of teacher assessment-test score gaps across fields, race by race, and gender by gender. Interestingly, the correlation is significant and above 0.5 for all races expect Pacific islanders. Moreover, the correlation is remarkably stable across races – from 0.445 to 0.552 –, 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.

4.6 Parental perceptions when matched to a teacher of their own race

Teachers may capture skills that are not captured in test scores. Stereotype threat is one identification issue that has been partly addressed by adding a child fixed effect. This fixed effect however fails to control for stereotype threats when it depends on the race of the teacher. If children perform better in the classroom than in the tests when matched to a teacher of their own race, my identification strategy will capture a mixture of student behavior rather than teacher behavior. However the dataset contains parental perceptions of students' skills and attentiveness. A potential falsification check is therefore to estimate the effect of same race assessment on parental perceptions rather than teacher perceptions. Any significant effect of same race assessment on parental perceptions would then reject the grading discrimination story.

Parents answer the following questions

- Question 1: Does your child learn, think and solve problems better, as well, slightly less or much less than other children of his age?
- Question 2: Does your child pay attention better, as well, slightly less or much less than other children of his age?

That is, four answers for each question. These questions are much less precise and do not cover the same scope as the teacher assessments. Under the maintained assumptions, parental perceptions should not depend on same race assessment conditionally on children effects and teacher effects.

$$\begin{aligned} p_{q,i,t} &= Pr[\text{Thinks Better}_{q,i,t} = 1 | \text{Same Race}_{i,f,t}, i, J(i,f,t), y_{i,f,t}] \\ &= \Lambda(\mu_{J(i,f,t)} + \delta y_{i,f,t} + u_{i,f} + \alpha_r \text{Same Race}_{i,f,t}) \end{aligned} \tag{17}$$

With the same specification for all four answers of the two questions. $p_{q,i,t}$ is the probability of answering that child i thinks better than children of his age, in period t for question $q = 1, 2$. Other notations are as before. Since it is not possible to identify such a specification with both the teacher and the child fixed effects, I only identify it with one of the effects. The identification of such a specification is described in Chamberlain (1980).

Results presented in table 22 indicate that parental perceptions do not indeed significantly depend on whether the child is matched to a teacher of the same race. They do however vary with the test score in English, suggesting that parents are at least partly aware of their child's skills. A 1 S.D. increase in test scores increases the relative probability of parents declaring the child as better able to learn, think and solve by 3.3% of a standard deviation.

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 reflect the ranking of children's cognitive skills.

I computed the child's rank in test scores and teacher assessments within surveyed children in the classroom. The small number of surveyed children per teacher is not an issue given that teachers fill assessment questionnaires only for the surveyed ones.

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 the same gender.

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

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

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

Results are presented in table 20. OLS estimates of same race teachers are between 0.09 ranks (Mathematics) and 0.119 ranks (English). Controlling for child fixed effects, this effect falls and is only significant 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.

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.

Relative ranking might be the whole story ...

In fact, relative ranking could potentially explain my main results. I design a small model that explains that identification issue and I test the hypothesis in the dataset. The results do not support ranking as the main explanation for the results.

Assume teachers order students on a rigid scale, and the absolute value of the assessments does not matter to them. Blacks could be overassessed whenever (i) they are more likely to be matched to other black kids than white kids (ii) black kids have on average lower test scores. If black students are more likely to be matched to underachievers when matched to a teacher of the same race than when not, then the effect of same race assessments might just reflect ranking and not teachers' perceptions.

I will 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. The 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.

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

For the sake of clarity, 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 a strong assumption but 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 $p = E(r' = b|r = b)$ that doesn't change with test scores, $r' \perp y|r$. Then,

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

And, symmetrically for whites:

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

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

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

If white children have uniformly better test scores and if there is some degree of de jure 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.

So, is relative ranking the whole story?

For this to explain our main result one would require that 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, conditionally on child and teacher fixed effects. In this case, peers' test scores would be correlated with same race teacher conditionally on the covariates and this invalidate the causal interpretation of the identification strategy.

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

Table 23 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 taught by a teacher of the same race. Interestingly this effect disappears in column (2), where I control for a child fixed effect in a conditional logit regression. That is, looking at the same child moving from a teacher of the same race to a teacher of a different 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 24 is another piece of evidence that suggests relative ranking is not the whole story. It is the results of the baseline regression of table 8 with an additional control for peers' average test score. These two tables are very similar, and the hypothesis that the coefficients of interest (column (8)) are equal between those two tables cannot be rejected at 95%. Controlling for peers' average test scores, child effects, teacher effects, and the test score, being assessed by a teacher of the same race increases test scores by 7.2% of a SD in mathematics and 4.4% of a SD in English.

6 Conclusion

This paper uses a unique US longitudinal dataset 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 race, but not of their gender. This effect is mainly due to the better grades given to non-hispanic black children by non-hispanic black teachers and to hispanic children by hispanic teachers. It should be noted that a conservative interpretation of the results cannot tell whether black teachers overassess black pupils or whether white teacher underassess black pupils. The same reasoning applies to hispanic, any race pupils.

Even if the absolute value of assessments is affected, the ranking of children within classrooms is not significantly affected. This is the first large scale analysis of teacher assessments vs. test scores that uses US elementary education data. Results highlight the fact that the teachers' races determine their perceptions of students' skills. Controlled experiments on teachers' perceptions in U.S. classrooms would be needed to assess both (i) how they affect children performance (ii) how public policies can change teachers' perceptions of their students.

References

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

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

Chamberlain, G. (1980), 'Analysis of covariance with qualitative data', *Review of Economic Studies* **47**(1), 225–38. available at http://ideas.repec.org/a/bla/restud/v47y1980i1p225-38.html.

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

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

Coate, S. & 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.

Fryer, R. G. & Levitt, S. D. (2004), 'Understanding the black-white test score gap in the first two years of school', *The Review of Economics and Statistics* **86**(2), 447–464.

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

Fryer, R. G. & 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.

Hanushek, E. A. (2003), 'The failure of input-based schooling policies', *Economic Journal* **113**(485), F64–F98.

Hanushek, E. A. & Kain, J. F. (2005), 'Teachers, schools, and academic achievement', *Econometrica* **73**(2), 417–458.

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. & 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.

Öster, A. (2006), Parental unemployment and children's school performance, Working Paper Series 2006:5, IFAU - Institute for Labour Market Policy Evaluation.

Pelcovitz, D., Kaplan, S., Goldenberg, B., Mandel, F., Lehane, J. & 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. (2007), 'The statistical theory of racism and sexism', *American Economic Review* **62**(4), 659–661.

Rose, L. C. & Gallup, A. M. (2006), *The 38th Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes Toward the Public Schools*, Phi Delta Kappa International.

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

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

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

Wood, G. H., Darling-Hammond, L., Neil, M. & Roschewski, P. (2007), Refocusing accountability: Using local performance assessments to enhance teaching and learning for higher order skills, Technical report, Briefing Paper Prepared for Members of The Congress of The United States, The National Center for Fair and Open Testing.

Appendix: The Effect of Same Race and Same Gender Teaching on Test Scores

The beneficial effects of same race or same gender teaching have been assessed by Dee in the NELS:88 and Project STAR data. The same analysis can be carried out in the ECLS-K with little effort. Moreover, the identification hypothesis can be relaxed. This section therefore replicates his results (Dee 2005*a*, Dee 2005*b*, Dee 2004), using the ECLS-K.

Teachers and The Gender Gaps in Student Achievement (Dee 2005)

Thomas Dee assesses the effect of the teacher's gender on test scores using the National Education Longitudinal Study of 1988 (NELS:88). It is a nationally representative, longitudinal study of 8th graders that began in 1988. Dee's analysis relies on comparisons within children and within teachers of the effect of the teachers' gender on outcomes.

I replicate his estimations using the ECLS-K. A few differences are interesting to notice. First, children are younger, they are followed from kindergarten to 5th grade whereas NELS:88 children are in 8th grade. Second, I rely on longitudinal identification which allows to have a different identification hypothesis, ie exogenous mobility.

The effect of same gender teaching relies on comparisons within teachers and within children.

$$y_{i,f,t} \quad = \quad \mu_{J(i,f,t)} + \gamma_g \text{Same Gender}_{i,f,t} + \lambda Z_{i,f,t} + u_i + \varepsilon_{i,f,t} \tag{20}$$

Notations are mine. $y_{i,f,t}$ is child i's outcome in field f = Maths or English, and t the time period, from fall kindergarten to spring 5th grade. $Z_{i,f,t}$ is a set of control such as time dummies.

The identification relies on the assumption of exogenous mobility. Mobility from/to same gender teachers should not be correlated with time varying characteristics that have an impact on test scores. If, for instance, children whose parents experience unemployment shocks – which may affect test scores (Öster 2006) – are more often matched with teachers of the same gender, the effect of same gender teaching will be underestimated. This is clear in the first differenced version of the specification.

$$\Delta y_{i,f,t} \quad = \quad \gamma_r \Delta \text{Same Gender}_{i,f,t} + \Delta \mu_{J(i,f,t)} + \lambda \Delta Z_{i,f,t} + \Delta \varepsilon_{i,f,t} \tag{21}$$

Unemployment shocks could appear in the residual $\Delta\varepsilon_{i,f,t}$ and be correlated with $\Delta \text{Same Gender}_{i,f,t}$. A similar robustness check that was performed in the body of the paper can also be performed here.

Results are presented in table 21. Interestingly, results are very similar to Dee (2005*b*)). The effect of same gender teaching on test scores is around 5.9% of a standard deviation in mathematics and 6.2% of a standard deviation in English. In mathematics, the comparison of OLS, child fixed effects, teacher fixed effects estimates (Columns (1) to (4)) indicates that child fixed effects are correlated with same gender teaching. In English, the comparison of estimates (Columns (5) to (8)) indicates that same gender teaching is correlated with child fixed effects.

Teachers, Race and Student Achievement in a Randomized Experiment (Dee 2004)

Dee 2004 uses Project STAR's experiment to estimate the effect of same race teaching on outcomes. It is a large scale randomized experiment from 1st to 3rd grade on the achievement benefits of small class sizes. Assignment of teachers and children to small classes was intended random. The design of Project STAR however alleviates the issue of nonrandom teacher child sorting, even though Hanushek formulated strong criticisms on the design of the experiment (Hanushek 2003).

The ECLS-K is not an experiment. A way to solve the problem of nonrandom child-teacher matching is to estimate the effect of same race teaching conditionally on child and teacher fixed effects. However here Project STAR and the ECLS cover the same part of the curriculum.

The specification is similar to (20). The same identification issues apply. Results are presented in table 21. In English (columns (5) to (8)), the effect of same race teaching is around 5.3% conditionally on teacher and child fixed effects. The comparison of OLS estimates and child fixed effects estimates suggest that unobserved children traits are negatively correlated with same race teaching, ie unobserved child ability may be correlated with being matched to a same race teacher. In mathematics, the effect of same race teaching is not significant, neither in OLS nor conditionally on teacher and child fixed effects. This is in stark contrast with Dee (2004).

Data Appendix

The Early Childhood Longitudinal Study follows a representative sample of around 20,000 children of the 1998-1999 kindergarten cohort. Data has been collected at six points in time: fall kindergarten, spring kindergarten, fall and spring grade 1, spring grade 3 and spring grade 5. Only a subsample was interviewed in fall grade 1 therefore I only consider the remaining five points of data collection. All of my regressions are weighted according to the instructions of the data designers. I describe below how variables were coded.

Teacher Assessments and Test Scores

Teacher Assessments (variables T1ARSLIT to T6ARSLIT in English, T1ARSMAT to T6ARSMAT in Mathematics) are taken from the indirect cognitive assessments or *Academic Rating Scale*. They are standardized year by year with a mean of 50 and a standard deviation of 10, before the exclusion of missing data.

Test scores (variables C1RRTSCO to C4RRTSCO and C5R2RTSC, C6R3RTSC) are extracted from direct cognitive assessments data. Test scores are standardized year by year with a mean of 50 and a standard deviation of 10 before the exclusion of missing data.

Teachers' and Students' races and ethnicities

Children's race & ethnicity (variable RACE) was reported by the parents in one question. Thus 'White, Non Hispanic' is an item that combines both race and ethnicity. This is not the case with teachers' races (variables B1RACE1-B1RACE5 to B5RACE1-B5RACE5 and J61RACE1-J61RACE5, J62RACE1-J62RACE2 for race, B1HISP to B5HISP and J61HISP-J62HISP for ethnicity).

I combined race and ethnicity to make children's and teachers' race comparable. Therefore in the paper, 'same race' should be understood as same race, same ethnicity.

Weights

Unless mentioned in the tables, regressions are weighted to correct for movers' subsampling. In order to control costs, the National Centre for Education Statistics only followed a random subsample of movers. I therefore use weights, variables `C1WEIGHT` to `C6WEIGHT`. Longitudinal weights were not available for children from kindergarten to grade five.

Finding programs on the web

My programs for the estimation of two way fixed effects models are available on the website of the Cornell Institute for Social and Economic Research, at http://repository.ciser.cornell.edu/viewcvs-public/cg2/branches/stata/ Other STATA do files can be found on the web, at http://www.pse.ens.fr/junior/ouazad/.

	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)

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.

Source: Early Childhood Longitudinal Study, Kindergarten Cohort of 1998/1999.

Table 2: Descriptive Statistics – All Periods Pooled

	— All Teachers —				- English Teachers -	- Mathematics Teachers -
	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

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 3: Racial and Gender Diversity among Teachers From Kindergarten to Grade 5

		Teacher's Race		
	Mean	Same Race	Different Race	*Difference*
	(1)	(2)	(3)	*(4)=(2)-(3)*
Mathematics Test Scores				
Test Score	50.245	51.503	48.209	*3.293***
	(9.878)	(9.629)	(9.936)	*[0.092]*
Teacher Assessments	50.210	51.030	48.883	*2.147***
	(9.917)	(9.779)	(9.994)	*[0.093]*
Teacher Assessments - Test Scores				
... Indian Child	1.259	4.832	0.569	*4.263***
	(8.818)	(8.634)	(8.691)	*[0.796]*
... Asian Child	-0.703	-1.489	-0.629	*-0.860*
	(9.470)	(11.423)	(9.264)	*[0.653]*
... Black, African American Child	2.109	2.665	1.891	*0.774***
	(8.924)	(9.173)	(8.816)	*[0.248]*
... Pacific Islander Child	0.666	2.367	0.501	*1.866*
	(8.032)	(6.728)	(8.134)	*[1.165]*
... 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]*

Source: Early Childhood Longitudinal Study, Kindergarten Cohort of 1998/1999.

Table 4: Descriptive Statistics – Test Scores, Teacher Assessments, Racial Matching – Mathematics

		Teacher's Race		
	Mean	Same Race	Different Race	*Difference*
	(1)	(2)	(3)	*(4)=(2)-(3)*
English Test Scores				
Test Score	50.332	51.340	48.700	*2.641*****
	(9.762)	(9.522)	(9.923)	*[0.076]*
Teacher Assessments	50.381	51.181	49.084	*2.097*****
	(9.848)	(9.711)	(9.930)	*[0.077]*
Teacher Assessments - Test Scores				
... Indian Child	2.527	5.332	2.010	*3.322*****
	(8.200)	(7.791)	(8.172)	*[0.642]*
... Asian Child	-0.774	-0.713	-0.781	*0.067*
	(8.435)	(8.387)	(8.442)	*[0.462]*
... Black, African American Child	1.421	2.543	0.961	*1.583*****
	(8.145)	(8.344)	(8.018)	*[0.184]*
... Pacific Islander Child	0.052	3.616	-0.333	*3.950*****
	(7.849)	(8.155)	(7.723)	*[0.926]*
... 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]*

Source: Early Childhood Longitudinal Study, Kindergarten Cohort of 1998/1999.

Table 5: Descriptive Statistics – Test Scores, Teacher Assessments, Racial Matching – English

		Teacher's Gender		
	Mean	Same Gender	Different Gender	*Difference*
	(1)	(2)	(3)	*(4)=(2)-(3)*
Mathematics Test Scores				
Test Score	50.245	50.088	50.387	*-0.300*****
	(9.878)	(9.477)	(10.227)	*[0.090]*
Teacher Assessments	50.210	50.622	49.836	*0.786*****
	(9.917)	(9.735)	(10.065)	*[0.090]*
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]*

Source: Early Childhood Longitudinal Study, Kindergarten Cohort of 1998/1999.

Table 6: Descriptive Statistics – Test Scores, Teacher Assessments, Gender Matching – Mathematics

		Teacher's Gender		
	Mean	Same Gender	Different Gender	*Difference*
	(1)	(2)	(3)	*(4)=(2)-(3)*
English Test Scores				
Test Score	50.332	51.199	49.538	*1.661*****
	(9.762)	(9.444)	(9.978)	*[0.075]*
Teacher Assessments	50.381	51.524	49.334	*2.190*****
	(9.848)	(9.748)	(9.823)	*[0.075]*
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]*

Source: Early Childhood Longitudinal Study, Kindergarten Cohort of 1998/1999.

Table 7: Descriptive Statistics – Test Scores, Teacher Assessments, Gender Matching – English

	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)	(0.137)	(0.093)	(0.113)	(0.094)	(0.163)
Test Score	0.591**	0.263**	0.588**	0.241**	0.659**	0.316**	0.669**	0.313**
	(0.004)	(0.009)	(0.004)	(0.011)	(0.003)	(0.006)	(0.003)	(0.006)
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)	(0.141)	(0.121)	(0.135)	(0.122)	(0.197)
Test Score	0.591**	0.262**	0.587**	0.241**	0.659**	0.316**	0.668**	0.314**
	(0.004)	(0.009)	(0.004)	(0.018)	(0.003)	(0.006)	(0.003)	(0.010)
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			

Reading: 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 8: Do Same Race or Same Gender Teachers Give Better Assessments Conditionally on Test Scores ?

	— Mathematics Teacher Assessment —					— English Teacher Assessment —				
	Fall Kindergarten	Spring Kindergarten	Spring Grade 1	Spring Grade 3	Spring Grade 5	Fall Kindergarten	Spring Kindergarten	Spring Grade 1	Spring Grade 3	Spring Grade 5
Boy	-1.131**	-1.229**	-0.294*	0.035	0.006	-1.796**	-2.377**	-2.413**	-2.591**	-3.208**
	(0.160)	(0.143)	(0.165)	(0.191)	(0.276)	(0.145)	(0.143)	(0.163)	(0.187)	(0.191)
Black, African American	-4.761**	-4.240**	-4.731**	-3.239**	-4.766**	-4.107**	-3.392**	-3.680**	-4.438**	-4.005**
	(0.227)	(0.205)	(0.246)	(0.292)	(0.458)	(0.205)	(0.204)	(0.244)	(0.286)	(0.312)
Hispanic, Any Race	-5.647**	-4.589**	-3.401**	-1.372**	-2.186**	-6.248**	-4.499**	-3.004**	-2.314**	-2.424**
	(0.210)	(0.194)	(0.226)	(0.267)	(0.359)	(0.192)	(0.193)	(0.224)	(0.262)	(0.251)
Asian	-1.885**	-1.640**	-0.959**	2.352**	2.356**	-3.106**	-1.796**	-0.282	0.891**	1.613**
	(0.454)	(0.351)	(0.370)	(0.448)	(0.556)	(0.407)	(0.350)	(0.367)	(0.439)	(0.383)
Pacific Islander	-4.695**	-3.980**	-5.405**	-1.639*	-0.703	-5.078**	-2.852**	-4.800**	-2.496**	-2.803**
	(1.074)	(0.859)	(0.837)	(0.995)	(1.294)	(0.930)	(0.858)	(0.836)	(0.973)	(0.946)
Indian	-5.824**	-6.823**	-5.306**	-4.859**	-6.055**	-5.906**	-6.233**	-5.385**	-5.231**	-5.005**
	(0.604)	(0.556)	(0.660)	(0.777)	(0.990)	(0.543)	(0.550)	(0.652)	(0.755)	(0.717)
Observations	14,462	18,744	14,425	11,190	5,261	17,688	18,908	14,577	11,357	10,720
R Squared	0.07	0.05	0.04	0.02	0.04	0.08	0.05	0.04	0.05	0.05
F Statistic	117.41	108.99	70.23	24.93	26.19	161.11	114.87	77.68	61.94	68.39

Reading: 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 9: The Gaps in Teacher Assessments from Kindergarten to Grade 5

	Child Fixed Effect		Teacher Fixed Effect	
	(1)	(2)	(3)	(4)
Male	-1.914**	-0.747**	0.525**	0.528**
	(-1.914)	(-0.747)	(0.525)	(0.528)
Black, African American	-1.551**	-0.529**	-0.102**	-0.132**
	(-1.551)	(-0.529)	(-0.102)	(-0.132)
Hispanic, Any Race	-1.152**	-0.609**	-0.128**	-0.195**
	(-1.152)	(-0.609)	(-0.128)	(-0.195)
Asian	0.296**	-0.403**	-0.251**	-0.275**
	(0.296)	(-0.403)	(-0.251)	(-0.275)
Native Hawaiian, other Pacific Islander	-2.493**	-1.811**	0.246**	0.191**
	(-2.493)	(-1.811)	(0.246)	(0.191)
American Indian or Alaska Native	-2.550**	-1.666**	0.090**	0.075**
	(-2.550)	(-1.666)	(0.090)	(0.075)
Teacher's Tenure				-0.011**
				(-0.011)
Teacher's Experience				-0.004**
				(-0.004)
Child's behavior controls	No	Yes	-	-
Other controls	-	-	– Grade Dummies –	
F Statistic	77.41	667.63	4.08	4.69
R Squared	0.04	0.29	0.00	0.00
Number of Observations	21,409	21,409	6,093	6,093

Reading: Male children' fixed effects are 19.1% of a standard deviation lower than female children' fixed effects when not controlling for the child's behavior. Male teachers' fixed effects are 5.2% 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 10: The Analysis of Pupil Effects – The Analysis of Teacher Effects

	Mathematics Teacher Assessments									
	$\theta = 0.0$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.4$	$\theta = 0.5$	$\theta = 0.6$	$\theta = 0.7$	$\theta = 0.8$	$\theta = 0.9$
Same Race Teacher	0.711**	0.620*	0.506**	0.360	0.164	-0.111	-0.523**	-1.213**	-2.597**	-6.717**
	(0.251)	(0.356)	(0.167)	(0.329)	(0.245)	(0.224)	(0.219)	(0.224)	(0.287)	(0.554)
Corrected Test Score	0.241**	0.268**	0.301**	0.345**	0.402**	0.483**	0.605**	0.809**	1.216**	2.427**
	(0.011)	(0.008)	(0.008)	(0.016)	(0.023)	(0.017)	(0.031)	(0.049)	(0.047)	(0.148)
	English Teacher Assessments									
	$\theta = 0.0$	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.4$	$\theta = 0.5$	$\theta = 0.6$	$\theta = 0.7$	$\theta = 0.8$	$\theta = 0.9$
Same Race Teacher	0.435**	0.327**	0.193	0.021	-0.208	-0.525**	-0.997**	-1.763**	-3.197**	-6.404**
	(0.124)	(0.158)	(0.155)	(0.143)	(0.147)	(0.123)	(0.130)	(0.128)	(0.191)	(0.244)
Corrected Test Score	0.313**	0.348**	0.391**	0.446**	0.520**	0.622**	0.772**	1.016**	1.470**	2.461**
	(0.006)	(0.006)	(0.007)	(0.012)	(0.008)	(0.013)	(0.014)	(0.014)	(0.027)	(0.073)

Reading: 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 11: What Measurement Error in Test Scores Can Explain the Results?

		Race						Gender	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Correlation across fields of ...	All Children	Black	Non Hispanic White	Indian	Pacific Islander	Hispanic	Asian	Male	Female
Test Scores	0.739	0.745	0.709	0.731	0.674	0.722	0.737	0.740	0.753
Teacher Assessments	0.803	0.812	0.794	0.821	0.793	0.795	0.792	0.803	0.809
Δ = Test Scores - Teacher Assessments	0.532	0.530	0.528	0.506	0.445	0.526	0.531	0.512	0.552
Number of Observations				45,923					

Source: Early Childhood Longitudinal Study, Kindergarten Cohort of 1998/1999.

Table 12: Correlation of the Teacher Assessment–Test Score Gap Between Topics

Mobility Pattern	Count		Boy	White	Black, African American	Hispanic	Pacific Islander	Asian	Indian
00000	6759	(31.6 %)	0.48	0.04	0.25	0.35	0.03	0.16	0.04
00001	345	(1.6 %)	0.50	0.11	0.39	0.28	0.04	0.10	0.08
00010	283	(1.3 %)	0.49	0.14	0.39	0.35	0.02	0.09	0.02
00011	143	(0.7 %)	0.48	0.47	0.29	0.17	0.01	0.03	0.03
00100	580	(2.7 %)	0.51	0.25	0.33	0.27	0.01	0.08	0.06
00101	172	(0.8 %)	0.55	0.40	0.32	0.22	0.02	0.02	0.02
00110	169	(0.8 %)	0.51	0.52	0.18	0.28	0.00	0.01	0.01
00111	307	(1.4 %)	0.51	0.82	0.09	0.08	0.00	0.00	0.00
01000	260	(1.2 %)	0.54	0.68	0.19	0.10	0.00	0.02	0.00
01001	18	(0.1 %)	0.61	0.78	0.11	0.11	0.00	0.00	0.00
01010	30	(0.1 %)	0.50	0.83	0.13	0.00	0.00	0.03	0.00
01011	45	(0.2 %)	0.56	0.98	0.00	0.02	0.00	0.00	0.00
01100	155	(0.7 %)	0.48	0.86	0.08	0.05	0.00	0.01	0.00
01101	45	(0.2 %)	0.49	0.96	0.04	0.00	0.00	0.00	0.00
01110	116	(0.5 %)	0.51	0.91	0.05	0.04	0.00	0.00	0.00
01111	360	(1.7 %)	0.51	0.96	0.01	0.03	0.00	0.00	0.00
10000	653	(3.1 %)	0.55	0.75	0.11	0.11	0.00	0.01	0.01
10001	20	(0.1 %)	0.55	0.65	0.10	0.20	0.00	0.05	0.00
10010	29	(0.1 %)	0.55	0.66	0.14	0.10	0.00	0.10	0.00
10011	27	(0.1 %)	0.44	0.89	0.00	0.04	0.00	0.07	0.00
10100	26	(0.1 %)	0.42	0.73	0.00	0.23	0.04	0.00	0.00
10101	7	(0.0 %)	0.43	0.86	0.00	0.14	0.00	0.00	0.00
10110	18	(0.1 %)	0.50	0.72	0.06	0.22	0.00	0.00	0.00
10111	19	(0.1 %)	0.37	0.84	0.05	0.05	0.05	0.00	0.00
11000	2489	(11.6 %)	0.52	0.70	0.14	0.12	0.00	0.02	0.01
11001	236	(1.1 %)	0.57	0.63	0.15	0.18	0.00	0.03	0.01
11010	342	(1.6 %)	0.50	0.74	0.13	0.11	0.01	0.01	0.00
11011	473	(2.2 %)	0.52	0.91	0.03	0.05	0.00	0.00	0.00
11100	1626	(7.6 %)	0.52	0.81	0.06	0.10	0.00	0.02	0.01
11101	705	(3.3 %)	0.49	0.85	0.05	0.08	0.00	0.00	0.01
11110	1188	(5.5 %)	0.52	0.90	0.04	0.06	0.00	0.00	0.00
11111	3764	(17.6 %)	0.50	0.97	0.02	0.01	0.00	0.00	0.00

Reading: ‘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 13: Mobility Patterns in Mathematics: Same Race Teacher vs. Not Same Race Teacher

	Same Race Teacher			
	(1)	(2)	(3)	(4)
Boy	-0.001	0.000		
	(0.003)	(0.003)		
Black, African American	-0.645**	-0.644**		
	(0.005)	(0.005)		
Hispanic, Any Race	-0.675**	-0.676**		
	(0.004)	(0.004)		
Asian	-0.829**	-0.831**		
	(0.008)	(0.008)		
Native Hawaiian, other Pacific Islander	-0.820**	-0.820**		
	(0.016)	(0.016)		
American Indian or Alaska Native	-0.734**	-0.733**		
	(0.012)	(0.012)		
Approaches to learning		0.000	-0.000	-0.000
		(0.000)	(0.000)	(0.000)
Self-control		-0.001*	-0.001	-0.000
		(0.000)	(0.000)	(0.001)
Interpersonal skills		0.000	0.001*	0.001
		(0.000)	(0.000)	(0.000)
Externalizing Problems Behavior		-0.001**	0.000	-0.000
		(0.000)	(0.000)	(0.001)
Internalizing Problems Behavior		-0.001**	-0.000	-0.000
		(0.000)	(0.000)	(0.000)
Child Fixed Effect	No	No	Yes	Yes
Teacher Fixed Effect	No	No	No	Yes
F Statistic	4,031.979	2,283.608	4.444	9.026
R Squared	0.522	0.522	0.822	0.889
Number of Observations	48,065	48,065	48,065	48,065

Reading: 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 14: The Matching of Teachers to Children – Race, Gender and Behavior

	Progress in Mathematics Teacher Assesments				Progress 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.
From Diff. Race Teacher	0.336	0.289	0.709**	0.803	0.517**	0.440**	0.381**	0.451**
to Same Race Teacher	(0.207)	(0.302)	(0.231)	(0.776)	(0.118)	(0.149)	(0.127)	(0.206)
Progress in Test Score	0.123**	0.073**	0.088**	0.029	0.240**	0.205**	0.225**	0.176**
	(0.011)	(0.018)	(0.011)	(0.025)	(0.006)	(0.008)	(0.007)	(0.010)
F Statistic	11.707	5.290	6.930	1.250	84.509	70.017	95.937	1.128
R Squared	0.011	0.525	0.435	0.830	0.037	0.253	0.298	0.510
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	22,073				44,471			

Reading: 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 15: Do Same Race or Same Gender Teachers Give Better Assessments? – Robustness check

	Mathematics Test Scores				English Test Scores			
	(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	-0.623**	-0.182	-1.279**	-0.151	-0.670**	-0.168	-1.135**	-0.060
	(0.113)	(0.103)	(0.119)	(0.214)	(0.090)	(0.088)	(0.092)	(0.110)
Teacher Assessment	0.542**	0.106**	0.589**	0.118**	0.619**	0.191**	0.641**	0.212**
	(0.004)	(0.004)	(0.004)	(0.006)	(0.003)	(0.003)	(0.003)	(0.006)
F Statistic	1,514.093	75.332	1,916.364	9.540	2,747.560	265.335	3,674.768	8.865
R Squared	0.398	0.865	0.536	0.894	0.460	0.814	0.563	0.843
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			

Reading: 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 16: Do Same Race or Same Gender Teachers Give Better Assessments? – Falsification check

	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.
Male Teacher - Male Student	-0.620	-0.144	0.763**	0.649**	-0.361	-0.600	0.355*	0.099
	(0.424)	(0.544)	(0.263)	(0.221)	(0.333)	(0.384)	(0.205)	(0.320)
Female Teacher - Female Student	0.591**	0.546	-0.663**	-0.485	-0.134	0.108	-0.617**	-0.356
	(0.285)	(0.373)	(0.211)	(0.373)	(0.227)	(0.255)	(0.169)	(0.267)
Test Score	0.591**	0.262**	0.587**	0.240**	0.659**	0.316**	0.668**	0.314**
	(0.004)	(0.009)	(0.004)	(0.015)	(0.003)	(0.006)	(0.003)	(0.009)
F Statistic	1,163.014	75.003	1,561.925	4.150	2,385.942	263.030	3,241.784	5.601
R Squared	0.348	0.665	0.540	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			

Reading: 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 17: Do Same Gender Teachers Give Better Assessments? Analysis Per Gender

	Mathematics Teacher Assesments			
	(1)	(2)	(3)	(4)
	OLS	Child f.e.	Teacher f.e.	Two way f.e.
Same Race Teacher – White, Non Hispanic	0.318*	1.001**	-0.180	0.068
	(0.192)	(0.252)	(0.192)	(0.404)
Same Race Teacher – Black, African American	-0.479	0.325	0.568**	0.944**
	(0.337)	(0.480)	(0.244)	(0.471)
Same Race Teacher – Hispanic, Any Race	0.915**	-0.395	2.010**	1.460**
	(0.366)	(0.527)	(0.225)	(0.649)
Same Race Teacher – Asian	0.423	0.594	0.085	0.549
	(0.721)	(1.001)	(0.667)	(1.440)
Same Race Teacher – Native Hawaiian, other Pacific Islander	0.589	1.587	0.788	1.265
	(1.557)	(2.113)	(1.488)	(1.735)
Same Race Teacher – American Indian or Alaska Native	1.104	3.369**	1.816**	4.171**
	(0.836)	(1.117)	(0.764)	(1.671)
Test Score	0.591**	0.263**	0.589**	0.241**
	(0.004)	(0.009)	(0.004)	(0.012)
F Statistic	984.640	58.975	1,255.387	4.155
R Squared	0.348	0.666	0.540	0.787
Child Controls	Yes	No	Yes	No
Teacher Controls	Yes	Yes	No	No
Other Controls	— Time dummies —			
Number of Observations	48,065			

Reading: 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 18: Do Same Race Teachers Give Better Assessments? Analysis Per Race - Mathematics

	English Teacher Assesments			
	(1)	(2)	(3)	(4)
	OLS	Child f.e.	Teacher f.e.	Two way f.e.
Same Race Teacher – White, Non Hispanic	0.236	0.598**	-0.692**	-0.166
	(0.151)	(0.174)	(0.147)	(0.293)
Same Race Teacher – Black, African American	0.147	0.278	1.340**	1.129**
	(0.260)	(0.331)	(0.183)	(0.314)
Same Race Teacher – Hispanic, Any Race	1.465**	0.204	2.511**	1.202**
	(0.294)	(0.376)	(0.200)	(0.317)
Same Race Teacher – Asian	0.515	-0.676	-0.121	-1.193*
	(0.522)	(0.648)	(0.496)	(0.683)
Same Race Teacher – Native Hawaiian, other Pacific Islander	1.746	0.585	1.402	1.014
	(1.165)	(1.411)	(1.103)	(1.106)
Same Race Teacher – American Indian or Alaska Native	-0.156	0.106	2.079**	2.519**
	(0.655)	(0.785)	(0.606)	(0.845)
Test Score	0.659**	0.316**	0.671**	0.314**
	(0.003)	(0.006)	(0.003)	(0.009)
F Statistic	2,021.104	202.052	2,614.115	5.608
R Squared	0.437	0.699	0.554	0.773
Child Controls	Yes	No	Yes	No
Teacher Controls	Yes	Yes	No	No
Other Controls	— Time dummies —			
Number of Observations	67,855			

Reading: 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 19: Do Same Race Teachers Give Better Assessments? Analysis Per Race - English

	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
	(0.032)	(0.044)	(0.036)	(0.070)	(0.025)	(0.031)	(0.027)	(0.037)
Rank in Test Scores	0.791**	0.632**	0.700**	0.458**	0.826**	0.652**	0.767**	0.525**
	(0.003)	(0.005)	(0.004)	(0.013)	(0.002)	(0.004)	(0.003)	(0.005)
F Statistic	3,989.915	2,503.930	2,877.095	6.127	7,675.977	4,731.055	6,395.074	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
	(0.041)	(0.051)	(0.046)	(0.071)	(0.032)	(0.037)	(0.035)	(0.034)
Rank in Test Scores	0.791**	0.633**	0.700**	0.458**	0.827**	0.652**	0.767**	0.525**
	(0.003)	(0.005)	(0.004)	(0.011)	(0.002)	(0.004)	(0.003)	(0.008)
F Statistic	3,988.925	2,503.727	2,877.067	6.127	7,672.291	4,730.472	6,393.026	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			

Reading: 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 20: Grading on a Curve – Do Same Race or Same Gender Teachers Give Better Assessments?

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mathematics Test Score				English Test Score			
	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	-0.175	-0.123	-0.728**	-0.092	0.292**	0.384**	-0.102	0.534**
	(0.115)	(0.079)	(0.123)	(0.099)	(0.106)	(0.077)	(0.112)	(0.066)
F Statistic	402.278	5.031	394.483	10.797	352.438	7.208	375.065	8.738
R Squared	0.115	0.842	0.269	0.869	0.086	0.799	0.230	0.827
Same Gender Teacher	0.956**	0.517**	0.960**	0.589**	1.208**	0.548**	1.242**	0.627**
	(0.129)	(0.081)	(0.139)	(0.098)	(0.127)	(0.088)	(0.135)	(0.125)
F Statistic	405.261	8.502	395.479	10.810	356.955	8.507	381.498	8.739
R Squared	0.115	0.842	0.270	0.869	0.087	0.799	0.231	0.827
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	62,140				74,808			

Reading: 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 21: Assessing the Effect of Same Race or Same Gender Teacher on Test Scores

	Does your child learn, think and solve problems					Does your child pay attention			
	Better	As well	As well	Slightly less	Much less	Better	As well	Slightly less	Much less
	than other children of his age?					*than other children of his age?*			
Same Race Teacher									
Point Estimate	-0.007	0.000	0.017	-0.086	0.322	-0.061	0.023	0.018	0.111
Standard Error	(0.055)	(0.049)	(0.023)	(0.094)	(0.249)	(0.059)	(0.049)	(0.078)	(0.176)
Odds Ratio	0.993	1.000	1.017	0.918	1.380	0.940	1.023	1.018	1.117
English Test Score									
Point Estimate	0.033**	-0.013**	-0.038**	-0.032**	-0.010	0.016**	-0.002	-0.019**	0.013
Standard Error	(0.003)	(0.003)	(0.001)	(0.005)	(0.013)	(0.003)	(0.003)	(0.004)	(0.009)
Odds Ratio	1.033**	0.987**	0.963**	0.969**	0.990	1.016**	0.998	0.981**	1.013
Chi Squared	11,893.43	21,937.09	24,412.08	2,905.37	301.66	8,501.03	23,216.25	4,227.75	580.64
Pseudo R Squared	0.460	0.557	0.346	0.379	0.350	0.407	0.568	0.370	0.331
Child Fixed Effect	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Teacher Fixed Effect	No	No	Yes	No	No	No	No	No	No
Other Controls	– Time Dummies –								
Effective Number of Observations	34,146	34,146	66,049	34,146	34,146	27,546	27,546	27,546	27,546
Number of Observations	– 67,855 –					– 67,855–			

Conditional logit regressions of dummies for parental perceptions of the child's skills and attentiveness. The number of actual observations differs from the total number of observations since children for whom parental perceptions do not vary over time do not contribute to the likelihood. Reading: a 10% of a SD increase of the test score increases the relative probability of parents declaring the child as better able to learn, think and solve by 3.3% (first column, odds ratio of the test score).

**: Significant at 1%. *: Significant at 5%.

Source: Early Childhood Longitudinal Study, Kindergarten Cohort of 1998/1999.

Table 22: Falsification Check - Parental Perceptions and Same Race Teacher

	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**
Standard Error	(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**
Standard Error	(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		

Reading: 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 23: Is the quality of the peers correlated with Same Race Teaching?

	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	0.273*	0.708**	0.667**	0.718*	0.394**	0.410**	0.603**	0.438**
	(0.118)	(0.162)	(0.119)	(0.288)	(0.092)	(0.112)	(0.094)	(0.057)
Test Score	0.595**	0.260**	0.589**	0.240**	0.679**	0.322**	0.681**	0.316**
	(0.004)	(0.009)	(0.004)	(0.010)	(0.003)	(0.006)	(0.003)	(0.007)
F Statistic	1,113.462	72.342	1,472.433	4.155	2,314.140	251.432	3,099.028	5.615
R Squared	0.348	0.666	0.540	0.787	0.440	0.700	0.556	0.773
Peers Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
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			

Reading: 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 24: Controlling for Peers' Quality