Mathematics Teachers' Bias Against the Mathematical Ability of Female, Black and Hispanic Students.

Copur-Gencturk, Y.

Robinson-Cimpian, J. P.

Lubienski, S. T.

Thacker, I.

Manuscript accepted for publication in: *Educational Researcher*.

Abstract

Teachers of mathematics play a significant role in students' perceptions of their mathematical ability and future career choices. Although a variety of studies have investigated teachers' biases, most of this work does not distinguish between teachers' accurate assessments of their students' academic ability and their implicit biases. In this randomized controlled study (N = 390), teachers of mathematics were asked to evaluate 18 mathematical solutions that varied according to three degrees of correctness (incorrect, partially correct, and correct solutions). Gender- and race-specific names were randomly assigned to a series of mathematical solutions, and teachers were asked to first evaluate the correctness of each solution and then estimate the mathematical ability of the student whose work they had evaluated. Teachers displayed no detectable bias when assessing the correctness of fictitious students' solutions; however, some gender- and race-associated biases were revealed in evaluations of partially correct and incorrect responses. When such biases occurred, non-White teachers' estimations of students' mathematical ability favored White students (both boys and girls) over students of color, whereas (primarily female) White teachers' estimations of students' mathematical ability favored boys over girls. Possible reasons for these results are considered, including that study design elements that might have influenced results, as well as the hypothesis that teachers from stereotyped groups may internalize the societal stereotypes that they are targeted by, making them more susceptible to bias that favors advantaged groups. These results suggest that interventions may be needed to address teachers' subtle gender- and race-related biases regarding students' mathematical ability.

Introduction

The academic achievement gap between White students and students of color and the underrepresentation of women and persons of color in STEM majors are two of the nation’s most persistent challenges. A shortage of skilled workers from diverse backgrounds to fill STEM-related positions (Camp, 1997) not only has damaging impacts on the economy (Langdon, McKittrick, Beede, Khan, & Doms, 2011) but also contributes to the current inequity in our society. However, despite the efforts in recent decades to increase the participation of women and persons of color in STEM fields (President’s Council, 2012), fewer women and persons of color enroll and persist in college STEM majors compared with White males (NCES, 2017; NSF, 2015).

Some have argued that the gap in STEM education can be explained by biological differences (Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000; Geary, Saults, Liu, & Hoard, 2000; Murray, Herrnstein, 1994). However, evidence suggests that while achievement gaps by SES are evident at the start of elementary school (Fryer & Levitt, 2004), gaps by gender and race (after accounting for SES) appear and develop while students are in elementary school (Cimpian, Lubienski, Timmer, Makowski, Miller, 2016; Husain & Millimet, 2009; Lubienski, McGraw, & Strutchens, 2004; Robinson & Lubienski, 2011). This evidence that, after adjusting for SES, race- and gender-based achievement gaps appear and expand only after schooling begins suggests that achievement differences are not genetic (Cimpian et al., 2016; Reardon, Robinson-Cimpian, & Weathers, 2015). Societal stereotypes that women and persons of color possess less innate raw talent could explain the underrepresentation of women and African Americans in STEM fields, where brilliance is seen as vital for success (Leslie, Cimpian, Meyer, & Freeland, 2015). Research indicates that schooling factors and cultural stereotypes seem to play a substantial role in students' academic outcomes. That is, schools seem to play a role in shaping gaps, and it is therefore important to examine potential school-related factors, including the possibility of teacher bias.

Students are particularly vulnerable to stereotyping in the classroom, where teachers' perceptions and actions have substantial consequences for students' academic achievement, self-perception, and educational and life trajectory (Benner & Graham, 2011; Farkas, 2003). Studies have shown that teachers' instructional decisions are shaped by their perceptions of their students' cognitive abilities (e.g., Clark & Peterson, 1986; Hoge & Coladarci, 1989). Furthermore, teachers' perceptions of individual students' abilities and their expectations for success may affect these students' achievement through a mechanism known as the self-fulfilling prophecy (e.g., Rosenthal & Jacobson, 1968). For example, a student whose ability is overestimated by a teacher might be willing to put more effort into learning. However, self-fulfilling prophecy effects may be smaller than studies suggest, as teacher perceptions can reflect actual differences in performance rather than teacher bias (Jussim & Harber, 2005). Taken together, investigating teachers' implicit biases against girls and students of color is vital when addressing inequity issues in education.

Race and Gender Bias

The literature on implicit social cognition often categorizes gender- and race-based biases as either explicit or implicit. Explicit biases are discriminatory attitudes and stereotyping behaviors that individuals are consciously aware of, are intentional, and are under the control of the individual. Implicit biases are those that individuals are unaware of, operate below the surface of consciousness, are out of the control of the individual (Bargh, 1994; Gawronski & Bodenhausen, 2006; Greenwald & Banaji, 1995; Strack & Deutsch, 2004) and appear under ambiguous situations in which lacking information might be inferred from signals, such as a race or gender being associated with a first name (e.g., Aigner & Cain 1977; Arrow, 1973; Bertrand, Chugh, & Mullainathan, 2005; Bertrand & Duflo, 2016; Dovidio, Gaertner, & Validzic, 1998; Greenwald & Banaji, 1995; Phelps, 1972). Although one could argue that teachers do not work in low-information or ambiguous situations and that their perceptions of their students' ability may arise from an accumulation of their daily experiences with students, teachers have relatively limited information regarding their new class of students at the beginning of an academic year, which could contribute to the emergence of teachers' implicit biases.

Empirical studies on teachers' explicit and implicit biases are limited. Research on explicit biases suggests that relatively low percentages of teachers explicitly express the belief that boys hold greater innate mathematical ability than girls (Copur-Gencturk, Thacker, Quinn, & Ebby, 2019) and that, compared to the general public, teachers report more positive explicit racial attitudes towards people of color (Quinn, 2017). However, explicit biases predict only a small number of educational disparities that occur in classrooms compared with implicit biases (Nosek & Smyth, 2011), which are sometimes unrelated to or even opposing explicit biases (Greenwald & Banaji, 1995).

Teachers' implicit biases are usually studied in observational settings in which teachers' implicit, unconscious biases are defined by their over- or underestimation of students' ability, as measured by the variance in teachers' assessments of their students' ability not explained by the students' performance on a direct assessment (Kilday, Kinzie, Mashburn, & Whittaker, 2012; Mashburn & Henry, 2004; McKown & Weinstein; 2002). However, this approach can be problematic because it fails to distinguish teachers' accuracy from their biases, as several validity issues are involved in researchers' use of imperfect, misaligned measures, as well as from random error. Specifically, teachers' perceptions are based on formal and informal evaluations of students' performance over time, which a single student test may not capture well.

Only a handful of studies have employed experimental methods to examine teachers' implicit biases, despite a large body of work in experimental psychology and other fields in which experimental methods have been used to capture people's implicit biases (for reviews, see Bertrand & Duflo, 2016; Nosek, Hawkins, & Frazier, 2011). One method used in the psychology literature is to measure implicit bias in terms of the differences in reaction time in people's automatic associations between concepts (e.g., girl–art, boy–mathematics; for example, see the Implicit Association Test [Greenwald, McGhee, & Schwartz, 1998; Nosek & Smyth, 2011] or Nosek, Hawkins, & Frazier, 2011 for a review of implicit bias measures). For example, Dutch teachers' implicit racial associations between Dutch–good and Turkish/Moroccan–bad indirectly predicted race gaps in their students' achievement as mediated by teachers' expectations, even though explicit racial attitudes were not significant predictors (van den Bergh, Denessen, Hornstra, Voeten, & Holland, 2010). In another study, Nurnberger, Nerb, Schmitz, Keller, and Sutterlin (2016) found that German preservice teachers' implicit associations of boys with mathematics and girls with language predicted their stereotypical tracking decisions of hypothetical boys into advanced-track mathematics and girls into lower tracks. However, despite the popularity of measuring implicit bias by using reaction time, the predictive validity of implicit association tests for racial associations have been called into question in a metanalysis demonstrating mixed findings regarding the predictive effects on a wide host of outcomes (Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013). Implicit bias studies have also tended to isolate gender or racial biases and do not explore race–gender intersectional biases. Further, implicit association studies do not relate to any specific instructional practice that teachers use regularly, and thus may not illuminate biases that arise in task-specific instructional situations.

Audit studies offer an alternate experimental method for capturing implicit bias that is considerably more task specific. Audit studies are common in the social psychology literature, wherein names are randomly assigned to objects of evaluation. For example, Moss-Racusin and colleagues (2012) found that science faculty rated an applicant for a laboratory manager position that was randomly assigned a male name as being more competent and hirable than the applicant randomly assigned a female name (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012). As one of the few studies conducted with actual teachers, Anderson-Clark, Green, and Henley (2008) randomly assigned common Black and White first names and Black and White racial information (e.g., Black name and White race, Black name and Black race) to a vignette describing a typical fifth-grade student and asked teachers to predict these students' motivation and achievement behaviors. The authors found a significant main effect for the race-associated names on teachers' evaluations but no effect for explicit statements about the students' race. In another study, Harber and colleagues (2012) examined differences in the feedback provided by teachers to one of four essays to which White, Hispanic, and Black student names had been randomly assigned. They found that White teachers' feedback to Hispanic students was more positive and less critical than their feedback to White students. This was also true of White teachers' feedback for Black students, but only on subjective aspects of writing (i.e., content, not mechanics), and only when those teachers felt unsupported by school administrators and fellow teachers. Harber and colleagues argue that such patterns suggest that White teachers may be less critical of Black students because of their self-image anxieties (e.g., not wanting to appear racist). They also hypothesize that White teachers' might hold lower expectations of the English skills of Hispanic students, which might explain teachers' more positive assessment of their writing.

However, these studies consider racial and gender biases in isolation and do not consider race–gender intersections. That is, even though teachers may hold negative implicit and explicit stereotypes about the intellectual ability of persons of color or the mathematical ability of females, very few studies have investigated the intersection of these two stereotypes. Individuals may experience discrimination that is specific to their personal intersection of race, class, and gender (e.g., discrimination specifically toward Black women; Crenshaw, 1989) and may be vulnerable to discrimination because the negative effects of stereotypes may be additive (sometimes termed *double discrimination*; Crenshaw, 1989; Heilman, Wallen, Fuchs, & Tamkins, 2004).

A related literature on the match between student and teacher demographics.

One might suspect that the existence of teacher biases depends on the race/ethnicity and gender of the teacher, and relatedly, that the "match" between a teacher's and student's demographics (race/ethinicity, gender) may mitigate any such biases and ultimately lead to improved student outcomes. Indeed, a growing body of studies using K-12 national and statewide data on students who are matched with a teacher of similar race/ethnicity or gender have investigated the impact of this matching on perceptions of student ability and overall effects on student achievement. We first discuss the reseach on student achievement, then on perceptions of student ability.

First, a recent review of matching studies found mixed results with respect to *student achievement* outcomes (Redding, 2019). Although Redding (2019) concludes that student-teacher race match may be beneficial for Black students overall (partly because of more consistent findings when behavior and disciplinary outcomes are examined), findings related to student achievement are rather mixed. For example, large-scale studies with national and state-level datasets found statistically significant relationships suggesting that Black K-12 students matched with Black teachers scored between about 0.20 SDs *lower* to about 0.15 SDs *higher* on standardized assessments in mathematics than when instructed by a non-Black teacher (Clotfelter et al., 2007; Dee, 2004; Egalite, Kisida, & Winters, 2015; Fryer & Levitt, 2004). Some research suggests the possibility of longer-term benefits to Black students (Gershenson, Hart, Lindsay, & Papageorge, 2017). Other studies have found non-significant effects (e.g., Buddin & Zamarro, 2009a; 2009b; 2009c). And the findings for Hispanic students tend to be even more inconsistent (Redding, 2019). A related line of evidence shows that gender-matching reveals somewhat mixed results when mathematics achievement is the outcome; some studies find small effects (e.g., 0.003-0.009 SDs for early elementary students [Clotfelter et al., 2007], 0.01 SDs for middle school students [Buddin & Zamarro, 2009b], and 0.05 SDs in high school [Buddin & Zamarro, 2009c]; see also, Ouazad, 2008), while others find no statistically significant or significant but negative effects of gender matching on mathematics achievement (Buddin & Zamarro, 2009a; Dee, 2007; Ehrenberg, Goldhaber, Brewer, 1994).

Second, this research has examined the impact of race and gender matching on teachers' *perceptions of their students' academic ability* (Ehrenberg, Goldhaber, & Brewer, 1994; McGrady & Reynolds 2013; Ouazad, 2014). The results concerning mathematical ability from racial- and ethnicity-matching studies tend to be driven by White teachers underrating non-White students compared with their ratings of White students; however, the results differ for Hispanic and Black students and depend on the measures used. For instance, Ouazad (2014) used a national sample of K-5 student data (ECLS-K data) and found that White teachers gave significantly lower assessments of their Hispanic students' mathematical proficiency compared with those of their White students, although the results were not significant for Black students after including fixed effects for student, teacher, and grade and adjusting for subject-specific test scores. Similarly, McGrady and Reynolds (2013) used a national sample of 10th-grade student data (ELS data) and found that White mathematics teachers were significantly more likely to indicate that their class was difficult for Black students (but not for their Hispanic students) and that their Hispanic students (but not their Black students) had fallen behind compared with their White students after adjusting for school- and student-level fixed effects. Additionally, Redding's (2019) review suggests a mixed set of race-matching findings with respect to teachers' ratings of students' academics, ranging from significantly negative findings (-0.14 SDs) to significantly positive findings (0.06 SDs) for Black students and no significant findings for Hispanic students. With respect to gender, some evidence suggests that elementary school female teachers are more likely than male teachers to underestimate the math ability of girls, suggesting a possible negative effect of gender matching (Robinson-Cimpian et al., 2014).

It is also worth noting that a recent study by Papageorge, Gershenson, and Kang (2019) investigated how 10th grade teachers' expectations for their students' educational attainment affected later college-going. That study found that teachers were on average overly optimistic about students' educational outcomes, and that this optimism may have a positive effect on students actually going to college. White teachers were less optimistic about their Black students compared with other combinations of student-teacher race. This study differs from ours in important ways (i.e., our experiment with fictitious students looking at ratings of math abilities vs. their quasi-experiment with real students looking at the effects of attainment expectations), but it illustrates—as indeed, this broader literature does—that the expectations teachers hold for students are multidimensional and complex.

Despite the mixed findings from this related literature on student-teacher demographic matching, it raises two important issues. First, any biases that can be examined aggregated across teacher demographics may vary depending on some demographics of the teachers; thus, we will explore our results by race.[1] Second, any race differentials should be interpreted in relation to the possibility of student-teacher race-matching effects.

Current Study

The present study makes four important contributions to our knowledge of teachers' implicit biases. First, we overcome the limitations of existing studies that ignore the broader and varied knowledge that teachers may have of the students in their classrooms, and instead we control the knowledge that teachers have about students by using fictitious students who are given names commonly associated with Black, Hispanic and White (non-Hispanic) males and females. Second, we assess teachers' evaluations of the correctness of student solutions separately from their evaluations of the students' mathematical ability based on the same responses. This approach allows us to examine the nature of the differences rather than conflate teachers' evaluations of correctness with their evaluations of ability. Third, we take an intersectional approach that goes beyond examinations of gender apart from race or ethnicity. Fourth, we created enough ambiguity in students' performance for teachers to draw on their biases, leaving room for us to detect potentially biased responses, as established by convention in experimental studies (e.g., Heilman et al., 2004; Moss-Racusin et al., 2012). We also tested the

[1] We would like to explore differences by teacher gender, but too few males precludes that possibily with any degree of reliability.

robustness of teachers' implicit biases by gathering data from teachers via a survey containing students' solutions with a variety of correctness levels. By creating a situation in which mathematics teachers' bias might occur in actual classroom settings, in this study, we explored the following research questions:

- *When examining problem solutions of fictitious students, do teachers' ratings of students' correctness and mathematical ability differ depending on the gender or race/ethnicity of the name assigned to the student?*
- *Do teachers' own race, gender, and educational backgrounds predict their implicit biases?*

Methods

Study Context

The data for this study came from mathematics teachers who participated in professional development activities provided by state-funded Mathematics and Science Partnership (MSP) programs in 2014–2017 in a Southern state. We partnered with a statewide MSP network, which provided the email addresses of teachers who taught mathematics in K-12 settings as well as background information on the teachers who ultimately completed our survey. We did not inquire about teachers' gender and race in the survey to avoid calling participants' attention to their own gender and race, which could have affected their evaluations of students' work and aroused suspicion regarding the real purpose of the study.

In the invitation e-mail for the survey, we created a deceptive story so that we could assess teachers' subtle biases more accurately. Specifically, teachers were told that we were in the final stage of selecting items for an assessment that would capture the features of middle school students' mathematical knowledge and skills that were most essential to predicting their mathematical growth. They were told they were participating in a study to help us identify the assessment items that best predicted students' mathematical growth and that their feedback would be used to finalize items for that instrument. Furthermore, teachers were told that their responses would remain anonymous. Again, to avoid raising suspicion, teachers were not told that their responses would later be linked to their demographic or educational background. All participants rated the same student work, which was assigned randomized combinations of student names associated with Black, Hispanic and White girls and boys (see Figure 1).

---Insert Figure 1 here---

Analytic Sample

We restricted our analyses to teachers who rated all the student work, who completed the survey in a reasonable amount of time, and for whom we had data on their race or ethnicity. Specifically, we expected that a teacher who read all the instructions, read and solved the problems, and rated students' work could not complete the survey in less than 8 minutes. Therefore, we set a minimum time of 8 minutes and excluded teachers who finished the survey in less than 8 minutes (*N* = 29, *M* = 6.38 minutes, *SD* = 1.46). We also excluded teachers who took longer than 180 minutes to complete the survey (*N* = 16, *M* = 2606.93 minutes, *SD* = 3593.16). Our rationale for this decision was that teachers who completed the survey in multiple sittings might not remember the initial instructions asking teachers to pay attention to students' names[2]. We also excluded 15 teachers whose race/ethnicity information was missing because we required teacher race/ethnicity information for our second research question. Those who were

[2] We reran the analyses, excluding respondents below the 5th percentile (i.e., those who took less than 7.82 minutes to complete the survey) and above the 95th percentile (i.e., those who took more than 115.47 minutes to complete the survey). The results were similar, indicating they were robust to these changes in the analytic sample.

excluded from the study were not statistically different from those included in the study in terms of gender [$\chi(1, N = 432) = 1.304$, $p = .253$], educational degree [$\chi(3, N = 423) = .608$, $p = .895$], and years of teaching experience [$M_{\text{difference}} = 1.74$, $SD = 1.3$, $t(413) = 1.32$, $p = 0.19$] . As shown in Tables S1 and S2 in *Supplementary Materials*, the results were similar for the full and analytical sample, as well as for Black and Hispanic girls and Black and Hispanic boys.

As presented in Table 1, 87.4% of the participating teachers were female and 65.4% were White. The years of teaching experience ranged from 1 to 35, with an average of 10.6 years (*SD* = 7.88). Approximately one-third of the teachers had a master's degree.

--Insert Table 1 here--

Name Selection

Because our findings were contingent on the selected names being associated with a targeted gender and race, we searched common names associated with certain gender and ethnic/racial groups. We searched existing research on biases (e.g., Bertrand, Chugh, & Mullainathan, 2004), websites with baby names (e.g., baby center.com), and lists of names perceived as "Whitest" and "Blackest" (ABC News, 2006; Levitt & Dubner, 2009) to identify the most popular Hispanic, Black, and White names for boys and girls. We also conducted Internet searches for famous personalities having the names we identified. After identifying a set of names, we surveyed teachers, preservice teachers, and teacher educators (N = 57) to identify which race or ethnicity and which gender came to mind when they saw these names. We finalized the names used in the study based on the responses we received (Table 2).

--Insert Table 2 here--

Survey Development

We created surveys using released National Assessment of Educational Progress (NAEP) mathematics problems. Specifically, we examined all available extended-response NAEP items from 2003 to 2013 and selected 13 items that seemed likely to prompt a range of correct and incorrect student responses. We created a booklet that included these 13 items and distributed the booklet to 29 middle-grade students. We decided, based on the pool of student work collected from the students, to use 3 of the 13 NAEP items in our study because of the quality and variety of responses provided by the students. We then selected two incorrect, two partially correct, and two correct responses to each problem, which resulted in 18 different solutions. When selecting these responses, we ensured that the handwriting and language used in the responses did not align with potential gender- or ethnicity-related stereotypes.

The eighteen names (from Table 2) were placed on the 18 solutions collected from the middle-school students (see Figure 1 for an example). We created all possible forms by rotating the race/ethnicity and gender within the solutions for the three different problems and difficulty levels (i.e., in each form, for each difficulty level, one Black, White and Hispanic boy and girl were assigned to each mathematics solution, which resulted in 24 different forms). We then field tested the survey with teachers, student teachers, and teacher educators to ensure that the newly produced student work looked authentic[3]. To frame our study, we told participants that we needed their help to validate an assessment designed to identify students' mathematical ability. To justify including the student names and ensuring that participating teachers would pay attention to them, we warned the teachers that some students had put their names or other

[3] During the piloting phase, the first and third author asked a few participants who had completed the survey to give their opinion of the survey. We did not specifically ask whether they were suspicious of the study; rather, we asked them what they thought of the survey. We have only anecdotal evidence; however, none of the participants we spoke with mentioned the possibility we were checking for biases.

identifying information on their work during the field testing of items, and we asked them to report if they noticed any information other than the students' first names.

Variables Used in This Study

Correctness. Teachers were asked to evaluate the mathematical soundness of each student solution based on a 10-point scale ranging from *absolutely nothing correct* to *fully mathematically sound*.

Mathematical Ability. After teachers rated the correctness of a given solution, they were also asked to estimate the mathematical ability of student based mathematical knowledge and insight reveal in the student's solution, using a 7-point scale ranging from *very low mahtematical ability* to *very high mathematical ability*. This method was used because teachers' instructional decisions and interactions with students are shaped by their perceptions of students' mathematical ability. Additionally, we used two different scales for ability and correctness to make sure that teachers do not transfer their scores from correctness to ability.

Student Gender and race. Each student solution was randomly assigned to a female or male name associated with being Black, White or Hispanic (Table 2). Dummy-coded variables were created for White girls and Black/Hispanic girls and boys, respectively. The results are the same for Black and Hispanic girls and Black and Hispanic boys, therefore, we combine these two groups (see *Tables S1 and S2 in Supplementary Materials*).

Display order. The order of the student solutions was randomized, and a variable was created indicating the order in which teachers evaluated the student work.

Teachers' background characteristics. Demographic information on the teachers included their gender, race (dummy-coded variables for White teachers and non-White teachers, respectively), educational degree (a dummy-coded variable indicating whether teachers had a master's degree or not), certification type (a dummy-coded variable indicating whether teachers were alternatively certified or not), and years of teaching experience.

Analytic Approach

We examined teachers' implicit biases in two ways. First, we examined whether teachers' evaluations of the correctness of students' solutions varied by the assigned gender and race. To do so, we predicted the dependent variable, teachers' evaluations of the correctness of a solution as being a function of the gender and race assigned to a given solution. The teacher, item, and item order were added as fixed effects. The standard errors for these and all subsequent analyses were adjusted appropriately to account for clustering of solutions within teachers. We performed separate analyses for the three different levels of correctness of students' solutions. In doing so, we were able to examine the robustness and sensitivity of the findings and identify where teachers' implicit biases were more prevalent. Second, we examined whether teachers' ratings of their students' abilities varied by the assigned gender and race. We regressed teachers' ratings of students' ability on the aforementioned variables along with their evaluation of the correctness of a given solution. Adjusting for teachers' evaluations of students' performance allowed us to examine whether teachers' estimations of students' mathematical ability reflected accurate predictions based on their own assessment of students' performance or biases in their expectations based on gender and race.

We also examined the extent to which teachers' background characteristics might predict their biases. We used two-level hierarchical linear models (HLMs) to evaluate whether teachers' gender or educational background was linked to their biases. We also examined the interaction between teachers' race and students' race to find whether teachers from different races showed similar levels of bias. The intercept was estimated as random, whereas all the slopes were estimated with fixed effects. The assigned students' gender, race, and teacher-centered correctness of solutions were added as Level 1 predictors along with the item and item order as fixed effects. Teachers' race was added as a predictor for the slopes of student race and gender indicators to examine the interaction effect. Teachers' other background variables were added as Level 2, along with teachers' average ratings of correctness. Specifically,

$$TchrAbilityRating_{it} = \beta_{0t} + \beta_{1t}Whitegirl_{it} + \beta_{2t}Minoritygirl_{it} + \beta_{3t}Minorityboy_{it} + \beta_{4t}correctness_{it} + \sum item + \sum itemorder + \varepsilon_{it},$$

$$\beta_{0t} = \gamma_{00} + \gamma_{01}Trace_t + \gamma_{02}Tgender_t + \gamma_{03}Tcertifcation_t + \gamma_{04}Tmasters_t + \gamma_{05}Tteachingexperience_t + \gamma_{06}Mean\_Correctness + \omega_{ot}$$

$$\beta_{kt} = \gamma_{k0+}\gamma_{k1}\,Trace_t,$$

where β_{kt} is the coefficient for Level 1 predictors (k = 1–3).

Results

Teachers' Ratings of the Correctness of Students' Solutions. Our results indicated that teachers' evaluations of the correctness of students' solutions did not differ based on the student's assigned gender, race or ethnicity (see Figure 2 for partially correct solutions). Specifically, teachers' ratings of correctness were not different for boys or girls, or White, Black or Hispanic students, or any pairwise comparisons (e.g., White males vs. Black or Hispanic girls). This pattern held for all three solution types examined: partially correct, incorrect and correct (see *Supplementary Materials* for more information on the results).

--Insert Figure 2 here--

Teachers' Ratings of Students' Mathematical Ability. We examined whether teachers' ratings of their students' abilities varied by students' assigned gender, race or ethnicity. We regressed their ratings of students' ability on student gender, race and ethnicity as well as on their evaluations of the correctness of a given solution while including item, item order, and teacher fixed effects. Adjusting for teachers' evaluations of correctness that did not differ by gender or race helped improve precision and allowed us to investigate whether teachers' estimations of students' mathematical ability reflected accurate predictions based on their own assessments of students' performance or whether they represented biased expectations based on gender and race. As we had initially hypothesized that the more ambiguous solutions would leave the most room for personal judgements, we found that teachers' implicit biases were revealed in partially correct student solutions after their ratings of the correctness of a given solution were taken into account (see Figure 3). For partially correct responses, teachers' ratings of White-sounding names were rated significantly higher than those of Black- and Hispanic-sounding names (Cohen's $d = 0.16$, $p < .01$). Solutions with White-sounding names were rated significantly higher than those with Black- and Hispanic-sounding names for both boys and girls, respectively ($d = 0.18$ and $d = 0.20$, respectively, both *p*s < .05 after Bonferroni correction for multiple comparisons).

---Insert Figure 3 here---

We also detected teachers' biases favoring male students' ability for incorrect responses. ($d = .10$, $p < .05$; see left panel of Figure 4). Although we did not initially expect to detect biases in the incorrect responses, teachers' ratings of students' ability were higher for incorrect responses than we anticipated (almost 3 on the 1-7 scale), leaving more room for subjectivity in teachers' ratings than we expected. Hence, it is reasonable that we saw some statistically significant differences among the incorrect solutions. Note, though, that the magnitude and number of differences among the incorrect solutions are both smaller than those found among the partially correct solutions, consistent with our expectations. Also as expected, no significant differences were found among the correct solutions.

---Insert Figure 4 here---

After finding that overall, teachers rated the mathematical ability of White-sounding names higher than those of Black- and Hispanic-sounding names, we disaggregated results by teachers' race. For partially correct responses, White teachers exhibited no detectable bias (see middle panel of Figure 3). Rather, as Figure 3 (right panel) illustrates, non-White teachers assigned higher ability ratings to White-sounding names (*d* = 0.27, p < .01), especially relative to their ratings of Black- and Hispanic-sounding girl names (*d*s > 0.33, *p*s < .05 after Bonferroni correction). It is important that both White and non-White teachers rated the ability of students with Black- and Hispanic-sounding names similarly, and so neither group of teachers exhibited more bias against these students per se. However, the non-White teachers rated White-sounding names higher than did the White teachers. This is shown in Figure 3, where the highest bars are for White-sounding names rated by non-White teachers. For incorrect solutions, White teachers exhibited bias against female students for incorrect responses (see the middle panel of Figure 4; *d* = 0.11, *p* < .05), whereas non-White teachers did not display a statistically significant bias against any group's ability when examining incorrect solutions. None of the pairwise comparisons were significant for either White or non-White teachers (see middle and right panels of Figure 4).

Finally, we wanted to test the interaction effect between teachers' race and the assigned student race as well as whether any of the teachers' background characteristics were linked to their ability ratings. As shown in Table 3, non-White teachers' ability ratings for White students (both boys and girls) were significantly higher than White teachers' ratings of White boys. Additionally, none of the additional teacher background characteristics was associated with teachers' ratings. Specifically, teacher's gender, certification status, education level, and teaching experience did not significantly predict teachers' assessments of students' mathematics ability for partially correct solutions (all *ps* > .05).

---Insert Table 3 here---

Discussion

Research on teachers' differential evaluations of the mathematical ability and performance of students of different races and genders has yielded mixed results (cf. Madon et al., 1998; Hinnant, O'Brien, & Ghazarian, 2009; Jamil, Larsen, & Hamre, 2018), partly because of the difficulty of distinguishing between when teachers make valid inferences about their students and when they are expressing biases. The present study is unique in investigating mathematics teachers' subtle biases toward students from stereotyped groups. In this experiment, we found that teachers did not show biases when they evaluated the mathematical soundness of a given student's work (i.e., student performance); however, non-White teachers showed biases when predicting students' mathematical ability (i.e., students' potential) from partially correct responses. On the same work, they made higher estimations of the mathematical ability of males or White students compared with the ability of female or Black/Hispanic students. Results of our study also indicate that whenever we saw significant pairwise differences, the lowest rated group was always non-White females. This is especially important given that students' perceptions of their academic ability are developed based on messages they receive from their social environment, especially those of their teachers and parents. These messages potentially contribute to their self-efficacy, self-competence, and decision to select a STEM career (Eccles & Wang, 2016; Kim, Sinatra, & Seyranian, 2018).

Among teachers of color, we detected bias favoring White students. These results seem counterintuitive, especially given that the literature on demographic matching generally finds that students of color, especially Black students, may benefit academically from having a teacher of the same race. The leading hypotheses in this body of literature explain that racial matching may lead to positive academic outcomes and evaluations of students because students and teachers of the same ethnicity share a cultural understanding that influences teachers' instructional decisions and facilitates student–teacher and parent–teacher relationships, or because students may respond more positively to the teacher's instruction and identify with the teacher if they share the same race, reducing the stereotype threat (see Redding, 2019, for a review). However, another body of literature on internalized racism (e.g., hooks, 2004; Speight, 2007; Williams & Williams-Morris, 2000) and internalized sexism (e.g., Bearman, Korobov, & Thorne, 2009; Szymanski & Kashubeck-West, 2008) has shown that oppressed groups sometimes accept and perpetuate negative racial and gender stereotypes, although the evidence for this hypothesis is mixed (cf. Harper, 2006). As such, teachers of color may be more critical of students of color because internalized stereotypes may manifest as lower expectations for students of their own race and consequently have a negative impact on student achievement.

We contend that despite the seemingly contrary findings between the literature on racial matching and internalized racism, the phenomena are not mutually exclusive. Teachers and students of the same race may share cultural understandings that help create culturally relevant instruction for students of color, which could lead to improvements in the academic performance of students of color. At the same time, teachers of color may have internalized stereotypes that they themselves have been subjected to throughout their lives and now hold the (possibly implicit) belief that White students may be mathematically more capable, which in turn could affect their expectations of students of color. In that case, the positive impact on student learning created by a same-race teaching environment could be negatively affected by these beliefs. Thus, we argue that students of color may benefit *even more* from a teacher of color who does *not* have internalized stereotypes. How these hypotheses (regarding the benefits of a shared cultural understanding, students' responsiveness to teachers of the same race, and internalized racism) fit together is an open empirical question that needs answering.

Another potential explanation may be related to our study design. Specifically, White teachers may be more averse to appearing racist and may devote more attention to hiding their biases (e.g., Crosby & Monin 2007; Dovidio & Gaertner, 1986; Plant & Devine, 1998). White teachers may be especially cautious about hiding their biases in experimental settings. Similarly, White teachers may be more concerned about maintaining their self-image as a nonracist person and may thus give higher ratings to students of color to protect their own image of themselves (Harber et al., 2012). In contrast, teachers from stigmatized groups may assume that they do not have biases; thus, they may be less cautious, which could have led us to capture only their biases in this study. Future studies could be designed to examine which factors might better explain teachers' implicit biases.

We also found White teachers favored male students compared with female students. Because most of the White teachers in our sample were female (91%), these teachers also may have internalized a stereotype to which they were exposed, causing them to perceive males as more mathematically capable than females. A study exploring bias in teachers' ratings of the mathematical ability of third-grade boys and girls in a nationally representative data set found that female teachers were more likely than male teachers to underestimate the mathematical ability of girls in comparison to similarly achieving and behaving boys, whereas male teachers demonstrated no evidence of bias either for or against girls (Robinson-Cimpian, Lubienski, Ganley, & Copur-Gencturk, 2014). Again, we cannot confidently explain why individuals from stigmatized groups would be likely to internalize societal stereotypes and exhibit biases that favored the dominant group. Therefore, future studies are needed to better understand the underlying mechanisms driving teachers’ implicit biases as well as how implicit bias captured in experimental studies is related to those captured in observational studies.

Our results indicated that biases were revealed when ambiguity in student solutions existed, allowing more teacher subjectivity. For correct responses, teachers did not exhibit any bias against students’ ability, whereas for partially correct responses, biases were identified. We also found biases for incorrect responses, on which teachers rated incorrect solutions higher than we anticipated, leaving room for more subjectivity. It is important to note that our findings are in the expected directions, that is, when there was a bias, it favored boys or White students. These results suggest that when the solutions were more ambiguous, there was more potential for bias against a group that was stereotyped as less mathematically or intellectually capable. These findings support theories of bias and discrimination which posit that biases are apparent in low-information situations (e.g., Aigner & Cain 1977; Arrow, 1973; Bertrand et al., 2005; Bertrand & Duflo, 2016; Dovidio et al., 1998; Greenwald & Banaji, 1995; Phelps, 1972). One might argue that teachers do not typically perform in low-information situations because they have ample opportunity to get to know their students through constant interaction during an academic year. However, we argue that, especially at the beginning of each academic year, teachers may make judgements in low-information situations and that such judgements could shape their perceptions of their students' abilities, which could lead to the mechanism identified in self-fulfilling prophecy studies (e.g., Rosenthal & Jacobson, 1966, 1968). However, it is also possible that when teachers get to know their students, they might change their expectations or evaluations of their students' ability. Still, it is important for teachers not to make initial judgements based on their perceptions of student race or gender; therefore, we argue that teacher education programs and professional development programs should be alerted to the potential for teachers to hold implicit biases.

It is important to note here that none of teachers' educational background indicators, such as years of teaching experience, certification type, or highest level of education completed, contributed to explaining the teachers' biases. Our results point to the importance of leveraging existing interventions or creating new, more targeted interventions that can prompt teachers to confront their biases in teacher education programs. For instance, Devine, Forscher, Austin, and Cox (2012) were able to create long-term reductions in adults' racial bias by using a 45-minute intervention intended to foster awareness of implicit bias, concern about the consequences of bias, and strategies that could be applied to reduce bias. Additional interventions have been created specifically to target the biases of mathematics instructors and faculty, for example, by guiding instructors to reflect on their biases through the lens of student participation patterns in their own classroom (Reinholz & Shah, 2018) and by developing equity-minded competencies while reflecting on achievement outcomes at their institution (Bensimon & Malcom, 2012).

Limitations

We must note some limitations of our study. First, because we collected data from a convenience sample, we do not know whether our findings would be generalizable to teachers across the United States. Future studies should examine whether the mathematics teacher population holds similar levels of implicit bias.

The second limitation of this study, and a limitation of audit studies in general, is the challenge of separating first names that signal race or social class associations. Although during our selection of the names, we tried to select names that were not associated with a certain socioeconomic status (SES), we did not collect data from teachers to ensure that the names we chose were not associated with a specific social class. Additionally, it is possible that a school's socioeconomic context could have affected the associations teachers make between names and student race, and more teachers of color may have taught in lower-SES schools than did White teachers. Hence, non-White teachers' apparent favoring of White students when rating students' ability could have been, at least in part, related to perceptions about students' social class. Given the high correlation between race and class in the United States, this possibility would still mean that White students' abilities are likely to be over-estimated by teachers of color, and that addressing issues of bias might require attention to both race- and class-related biases.

Additionally, although we tried to imitate potential areas where teachers would reveal their implicit biases in actual classroom settings (i.e., grading and evaluating student work), we could not rule out the fact that teachers might assess their own students' ability differently in real classroom settings. Nevertheless, it is important to note that teachers evaluate student work in relatively low-information situations early in a new academic year. As found in the "Pygmalion study," (Rosenthal & Jacobson, 1966; 1968) early impressions of students' potential could have an impact on student outcomes. Even so, more research is needed to check whether the bias captured in an experimental setting, as in this work, had any impact on student learning. Although some empirical evidence suggests that the implicit bias captured in experimental studies seems to be associated with the achievement of teachers' actual students (van den Bergh et al., 2010), more evidence is needed.

Finally, although participating teachers were informed that their role in this project was to help us identify items that might predict students' later mathematical success, and although we have anecdotal evidence from our pilot phase that suggested that respondents accepted the survey's rationale at face value, we did not systematically collect data that would reveal the extent to which the teachers in our sample believed this to be true. A similar study on writing feedback by Harber et al. (2012) asked respondents to rate their suspiciousness of the study's stated purpose, and the respondents reported generally low levels of suspicion and their suspicion was uncorrelated with any outcomes of interest in that study. Thus, although our study lacked a direct assessment of suspicion, this prior research and our anecodotal evidence suggest suspicion was likely low and not a primary driver of the results.

Conclusion

In conclusion, we used an experimental method to capture teachers' implicit biases in classroom settings by ruling out potential alternative explanations, such as those found in observational studies, and to help disentangle evaluations of correctness from impressions of ability. We hope this study will prompt present and future teachers to consider their own biases regarding the mathematics abilities of females and students of color. It would be helpful for teachers to know that they might hold deep-seated biases regarding students' abilities even while evaluating students' solutions fairly, regardless of their years of teaching experience and education and despite (or perhaps even because of) the societal biases they themselves face in mathematics.

Acknowledgments: We thank the volunteers, who are 2016 graduates of the middle school certification program in the College of Education at the University of Houston. We also thank Debra Plowman and David Quinn.

References

ABC News (2006). Top 20 'whitest' and 'blackest' names. *ABC News*. Retrieved from: https://abcnews.go.com/2020/top-20-whitest-blackest-names/story?id=2470131

Aigner, D. J., & Cain, G. G. (1977). Statistical theories of discrimination in labor markets. *ILR Review, 30*(2), 175-187.

Anderson-Clark, T. N., Green, R. J., & Henley, T. B. (2008). The relationship between first names and teacher expectations for achievement motivation. *Journal of Language and Social Psychology, 27*(1), 94-99.

Arrow, K. (1973). The theory of discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in labor markets* (pp. 3-33). Princeton, N.J.: Princeton University Press.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition.* Vol. 1: Basic processes (pp. 1–40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bearman, S., Korobov, N., & Thorne, A. (2009). The fabric of internalized sexism. *Journal of Integrated Social Sciences, 1*(1), 10-47.

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences, 107*(5), 1860–1863.

Benbow, C. P., Lubinski, D., Shea, D. L., & Eftekhari-Sanjani, H. (2000). Sex differences in mathematical reasoning ability at age 13: Their status 20 years later. *Psychological Science*, *11*(6), 474–480.

Benner, A. D., & Graham, S. (2011). Latino adolescents' experiences of discrimination across the first 2 years of high school: Correlates and influences on educational outcomes. *Child Development*, *82*(2), 508–519.

Bensimon, E. M., & Malcom, L. (2012*). Confronting equity issues on campus: Implementing the Equity Scorecard in theory and practice.* Sterling, VA: Stylus Publishing

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit discrimination. *American Economic Review, 95*(2), 94-98.

Bertrand, M. & Duflo, E. (2017). Field experiments on discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of economic field experiments*, Vol. 1 (pp. 309-393). Amsterdam: North Holland Publishing.

Betz, D. E., & Sekaquaptewa, D. (2012). My fair physicist? Feminine math and science role models demotivate young girls. *Social Psychological and Personality Science*, *3*, 738–746.

Buddin, R., & Zamarro, G. (2009a). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics, 66*, 103–115.

Buddin, R., & Zamarro, G. (2009b*). Teacher qualifications and middle school student achievement* (WR-671-IES). Santa Monica CA: RAND Corporation.

Buddin, R., & Zamarro, G. (2009c). *Teacher effectiveness in urban high schools* (WR-693-IES). Santa Monica, CA: RAND Corporation.

Camp, T. (1997). The incredible shrinking pipeline. *Communications of the ACM*, *40*(2), 103–110.

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. *AERA Open, 2*(4), 233285841667361.

Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 255–296). New York, NY: Macmillan.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and why do teacher creden- tials matter for student achievement?* (CALDER Working Paper 2). Washington, DC: Urban Institute.

Copur-Gencturk, Y., Thacker, I., Quinn, D., & Ebby, C. B. (2019, April). *K-8 Mathematics teachers' overall and gender-specific beliefs about mathematical aptitude.* Presentation given at the American Educational Research Association, Toronto, Canada.

Crenshaw, K. (1989). *Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics*. In D. K. Weisbert, Ed., *University of Chicago Legal Forum.* Philadelphia: Temple University Press.

Crenshaw, K. W. (1994). Mapping the margins: Intersectionality, identity, politics, and violence against women of color. In M. A. Fineman & R. Mykitiuk (Eds.), *The public nature of private violence* (pp. 93–118). New York: Routledge.

Crosby, J. R., & Monin, B. (2007). Failure to warn: How student race affects warnings of potential academic difficulty. *Journal of Experimental Social Psychology, 43*(4), 663-670.

Dee, T. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics, 86*, 195–210.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources, 42*(3), 528-554.

Dovidio, J. F., & Gaertner, S. L. (1986). *Prejudice, discrinzination, and racism.* Orlando, FL: Academic Press.

Dovidio, J. F., Gaertner, S. L., & Validzic, A. (1998). Intergroup bias: Status, Differentiation, and a common in-group identity. *Journal of Personality and Social Psychology, 75*(1), 109-120.

Eccles, J. S., & Wang, M. Te. (2016). What motivates females and males to pursue careers in mathematics and science? *International Journal of Behavioral Development*, *40*(2), 100–106.

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review, 45*, 44–52.

Ehrenberg, R. G., Goldhaber, D., & Brewer, D. J. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988. *Industrial and Labor Relations Review, 48*, 547–561.

Farkas, G. (2003). Cognitive skills and noncognitive traits and behaviors in stratification processes. *Annual Review of Sociology*, *29*(1), 541–562.

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.

Gawronski, B., & Bodenhausen,G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731.

Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, *77*(4), 337–353.

Gershenson, S., Hart, C. M. D., Lindsay, C. A., & Papageorge, N. W. (2017). *The long run impacts of same-race teachers match* (IZA DP No. 10630). Retrieved from http://ftp.iza.org/dp10630.pdf

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27.

Harber, K. D., Gorman, J. L., Gengaro, F. P., Butisingh, S., Tsang, W., & Ouellette, R. (2012). Students' race and teachers' social support affect the positive feedback bias in public schools. *Journal of Educational Psychology, 104*(4), 1149.

Harper, S. R. (2006). Peer support for African American male college achievement: Beyond internalized racism and the burden of acting White. *The Journal of Men's Studies, 14*(3), 337-358.

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: reactions to women who succeed at male gender-typed tasks. *Journal of Applied Psychology, 89*(3), 416.

Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*(3), 662.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297–313.

hooks, b. (2004). *We real cool: Black men and masculinity.* New York, NY: Routledge.

Husain, M., & Millimet, D. L. (2009). The mythical "boy crisis"? *Economics of Education Review*, *28*(1), 38–48.

Jamil, F. M., Larsen, R. A., & Hamre, B. K. (2018). Exploring longitudinal changes in teacher expectancy effects on children's mathematics achievement. *Journal for Research in Mathematics Education*, *49*(1), 57-90.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review*, *9*(2), 131–155.

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment*, *30*(2), 148–159.

Kim, A. Y., Sinatra, G. M., & Seyranian, V. (2018). Developing a STEM identity among young women: A social identity perspective. *Review of Educational Research*, 1–37.

Langdon, D., McKittrick, G., Beede, D., Khan, B., & Doms, M. (2011). *STEM: Good jobs now and for the future*. Washington, DC. Retrieved from http://www.esa.doc.gov/sites/default/files/stemfinalyjuly14_1.pdf

Lavy, V., & Sand, E. (2015). On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases (No. w20909). National Bureau of Economic Research.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347*(6219), 23–34.

Levitt, S. D., & Dubner, S. J. (2009). *Freakonomics: A rogue economist explores the hidden side of everything*. New York: HarperCollins.

Lubienski, S. T., McGraw, R., & Strutchens, M. (2004). NAEP findings regarding gender: Mathematics achievement, student affect, and learning practices. In P. Kloosterman & F.

K. Lester Jr. (Eds.), *Results and interpretations of the 1990 through 2000 mathematics assessments of the National Assessment of Educational Progress (pp. 305–336).* Reston, VA: National Council of Teachers of Mathematics.

Madon, S., Jussim, L., Keiper, S., Eccles, J., Smith, A., & Palumbo, P. (1998). The accuracy and power of sex, social class, and ethnic stereotypes: A naturalistic study in person perception. *Personality and Social Psychology Bulletin*, *24*(12), 1304–1318.

Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement: Issues and Practice, 23*(4), 16–30.

McGrady, P., & Reynolds, J. (2013). Racial mismatch in the classroom: Beyond Black–White differences. *Sociology of Education, 86*, 3–17.

McKown, C., & Weinstein, R. S. (2002). Modeling the role of child ethnicity and gender in children's differential response to teacher expectations. *Journal of Applied Social Psychology, 32*(1), 159–184.

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology*, *46*(3), 235–261.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474-16479.

Murray, C., & Herrnstein, R. (1994). *The bell curve: Intelligence and class structure in american life*. New York, NY: Free Press.

National Center for Science and Engineering Statistics (2017). *Women, minorities, and persons with disabilities in science and engineering: 2017*. Arlington, VA. Retrieved from www.nsf.gov/statistics/wmpd/

National Science Foundation (2015). *Science and engineering degrees, by race/ethnicity of recipients: 2002-12*. Arlington, VA. Retrieved from https://www.nsf.gov/statistics/2017/nsf17310/.

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences, 15*(4), 152–159.

Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal, 48*(5), 1125–1156.

Nürnberger, M., Nerb, J., Schmitz, F., Keller, J., & Sütterlin, S. (2016). Implicit gender stereotypes and essentialist beliefs predict preservice teachers' tracking recommendations. *The Journal of Experimental Education, 84*(1), 152-174.

Ouazad, A. (2014). Assessed by a teacher like me: Race and teacher assessments. *Education Finance and Policy, 9*, 334–372.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology, 105*(2), 171-192.

Papageorge, N. W., Gershenson, S., & Kang, K. M. (2019). Teacher expectations matter. *Review of Economics and Statistics.* Retrieved from: https://www.mitpressjournals.org/doi/abs/10.1162/rest_a_00838

Redding, C. (2019). A teacher like me: A review of the effect of student–teacher racial/ethnic matching on teacher perceptions of students and student academic and behavioral outcomes. *Review of Educational Research, 89*(4), 499–535.

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

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811.

President's Council of Advisors on Science and Technology. (2012). Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics, 130. Retrieved from http://www.esa.doc.gov/sites/default/files/stemfinalyjuly14_1.pdf.

Reardon, S. F., Robinson-Cimpian, J. P., & Weathers, E. S. (2015). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of Research in Education Finance and Policy* (pp. 1–29). Lawrence Erlbaum.

Reinholz, D. L., & Shah, N. (2018). Equity analytics: A methodological approach for quantifying participation patterns in mathematics classroom discourse. *Journal for Research in Mathematics Education*, *49*(2), 140-177.

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*(2), 268–302.

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers’ perceptions of students’ mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology*, *50*(4), 1262.

Rosenthal, R., & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Reports*, *19*(1), 115-118.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils’ intellectual development*. New York, NY: Holt, Rinehart and Winston.

Speight, S. L. (2007). Internalized racism: One more piece of the puzzle. *The Counseling Psychologist*, *35*(1), 126-134.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, *8*, 220–247.

Szymanski, D. M., Kashubeck-West, S., & Meyer, J. (2008). Internalized heterosexism: Measurement, psychosocial correlates, and research directions. *The Counseling Psychologist*, *36*(4), 525-574.

Williams, D. R., & Williams-Morris, R. (2000). Racism and mental health: The African American experience. *Ethnicity & Health*, *5*(3/4), 243-268.

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, *47*(2), 497-527.

Table 1. Descriptive statistics for teacher-level variables

Variable	*N*	Missing	*M*	*SD*	%
Teacher Gender		0			
Female	350				87.4
Male	40				12.6
Teacher Race/Ethnicity		0			
White	255				65.4
Black	48				12.3
Hispanic	68				17.4
Other	19				4.9
Teacher Educational Background		0			
Master's Degree or Higher (Yes)	120	0			30.8
Alternatively Certificated (Yes)	110	0			28.2
Teaching Experience (in Years)	377	13	10.6	7.88	

Note: *N* = 390 teachers.

Table 2. Names assigned to students' work

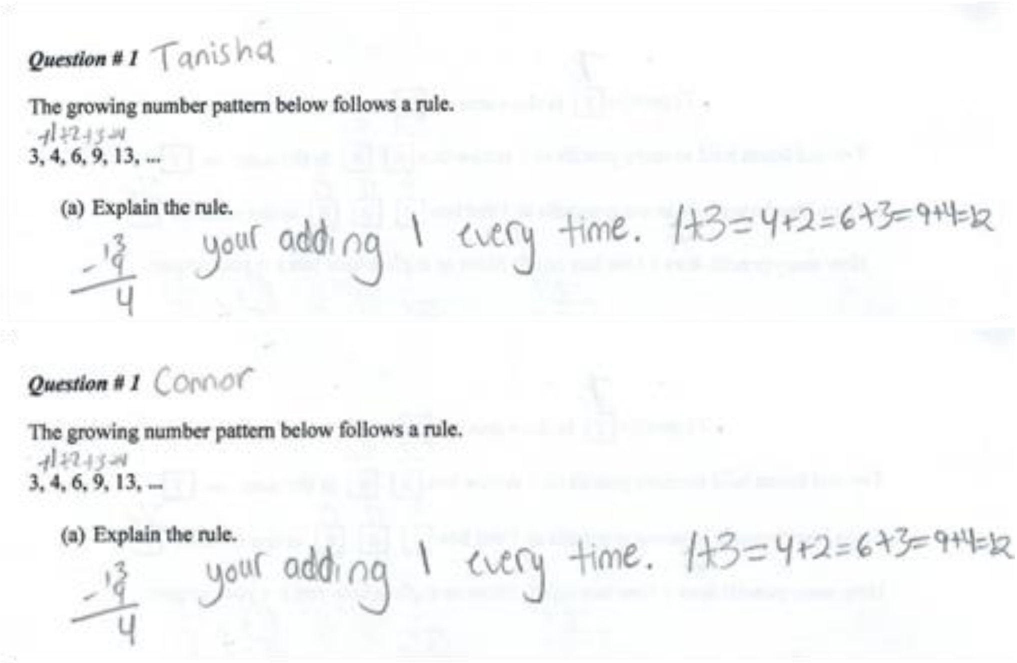
Gender	Black	White	Hispanic
Girl	Lakisha, Shanice, Tanisha	Emily, Katie, Molly	Blanca, Esmeralda, Rosalía
Boy	Tyrone, DeShawn, Trevon	Connor, Ethan, Todd	Alejandro, Diego, José

Table 3. Hierarchical linear model results for teachers' predictions of students' mathematical ability for partially correct responses.

Predictors	All Teachers
White boys x Non-White Teachers	**0.217****
	(0.082)
White girls xWhite Teachers	0.013
	(0.061)
White girls x Non-White Teachers	**0.227****
	(0.082)
Black and Hispanic boys x White Teachers	-0.047
	(0.053)
Black and Hispanic boys x Non-White Teachers	0.081
	(0.071)
Black and Hispanic girls x White Teachers	-0.065
	(0.053)
Black and Hispanic girls x Non-White Teachers	-0.017
	(0.071)
Teacher-centered correctness score	**0.488*****
	(0.009)
Teacher-level mean correctness score	**0.538*****
	(0.021)
Teachers' background characteristics	
Teacher gender (female)	0.042
	(0.074)
Alternatively certified (yes)	0.000
	(0.054)
Master's degree or higher (yes)	-0.002
	(0.049)
Teaching experience (in years)	-0.003
	(0.003)
Intercept	0.990***

	(0.199)
N (teacher)	390

**p* < .05, ***p* < .01; ****p* < .001 Significant statistics are in boldface. Number in parenthesis are standards errors.



Question # 1 Tanisha

The growing number pattern below follows a rule.

+1 +2 +3 +4

3, 4, 6, 9, 13, ...

(a) Explain the rule.

13
−9
4

your adding 1 every time. 1+3=4+2=6+3=9+4=12

Question # 1 Connor

The growing number pattern below follows a rule.

+1 +2 +3 +4

3, 4, 6, 9, 13, ...

(a) Explain the rule.

13
−9
4

your adding 1 every time. 1+3=4+2=6+3=9+4=12

Figure 1. The same student work assigned different names.

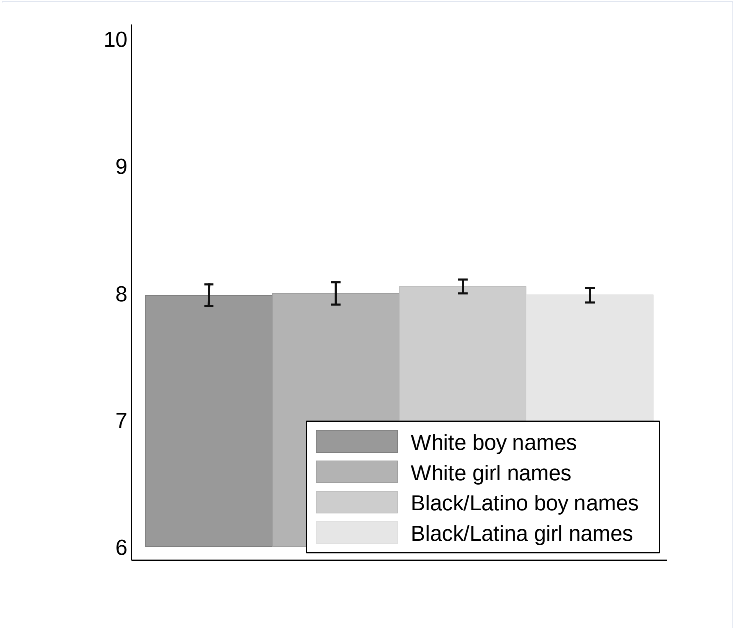


Figure 2. Mean correctness score for partially correct solutions, by teacher race and gender and race of student names. Heteroskedastic-robust standard errors clustered on teachers appear as bars around the mean estimate. Statistical significance (*5%, **1%) has been Bonferroni-adjusted for multiple subgroup comparisons. Models also include controls for item, item positioning on questionnaire, and teacher fixed effects.

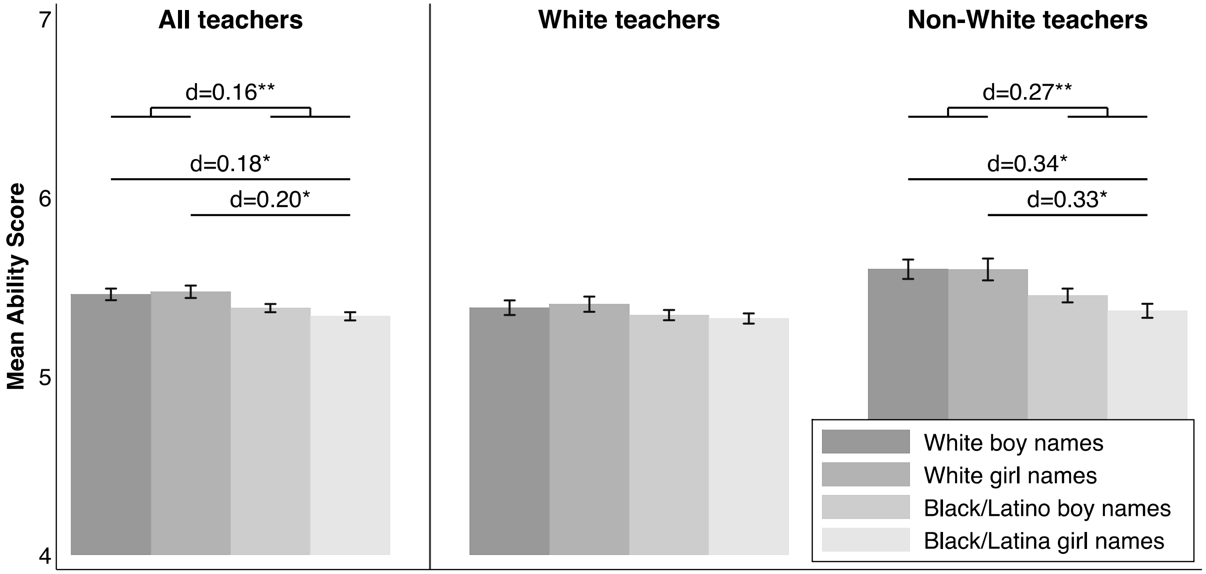


Figure 3. Mean ability score for partially correct solutions, by teacher race and gender and race of student names. Heteroskedastic-robust standard errors clustered on teachers appear as bars around the mean estimate. Statistical significance (*5%, **1%) has been Bonferroni-adjusted for multiple subgroup comparisons. Models also include controls for correctness ratings, item, item positioning on questionnaire, and teacher fixed effects.

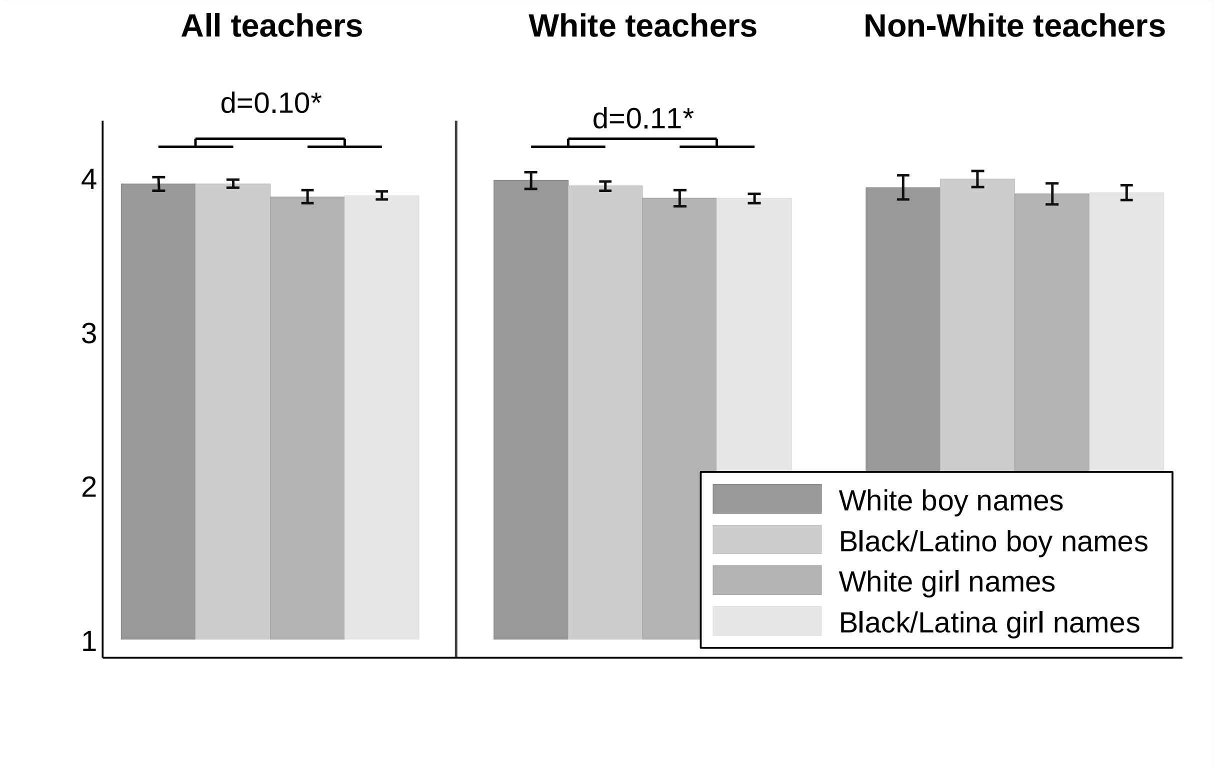


Figure 4. Mean ability score for incorrect responses, by teacher race and gender and race of student names. Heteroskedastic-robust standard errors clustered on teachers appear as bars around the mean estimate. Statistical significance (*5%, **1%) has been Bonferroni-adjusted for multiple subgroup comparisons. Models also include controls for correctness ratings, item, item positioning on questionnaire, and teacher fixed effects.

Supplementary Materials

Table S1. Means for teachers’ estimations of the correctness of students’ solutions

	Incorrect Responses				Partially Correct Responses				Correct Responses			
	Full Sample		Analytic Sample		Full Sample		Analytic Sample		Full Sample		Analytic Sample	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
White boys	3.57	0.10	3.46	0.11	7.93	0.08	7.98	0.08	8.28	0.10	8.36	0.10
White girls	3.66	0.10	3.56	0.11	8.03	0.08	7.99	0.09	8.21	0.09	8.28	0.10
Black girls	3.76	0.10	3.73	0.10	8.01	0.08	8.01	0.09	8.19	0.10	8.18	0.11
Hispanic girls	3.71	0.10	3.69	0.11	7.97	0.09	7.95	0.09	8.38	0.09	8.34	0.10
Black boys	3.69	0.10	3.61	0.10	8.14	0.08	8.12	0.09	8.29	0.09	8.35	0.10
Hispanic boys	3.73	0.10	3.62	0.10	7.96	0.08	7.97	0.09	8.28	0.10	8.35	0.10

Note: *N* = 450 teachers in the full sample and 390 teachers in the analytic sample. Mean scores reported here are adjusted for item, item positioning on questionnaire, and teacher fixed effects. Mean scores can range from 1 (*absolutely nothing correct) to 10 (fully mathematically sound.*

Table S2. Means for teachers' estimations of students' mathematical ability

	Incorrect Solutions				Partially Correct Solutions				Correct Solutions			
	Full Sample		Analytic Sample		Full Sample		Analytic Sample		Full Sample		Analytic Sample	
	Mean (SD)	Mean$_a$ (SD)	Mean (SD)	Mean$_a$ (SD)	Mean (SD)	Mean$_a$ (SD)	Mean (SD)	Mean$_a$ (SD)	Mean (SD)	Mean$_a$ (SD)	Mean (SD)	Mean$_a$ (SD)
White boys	2.93	2.99	2.89	2.96	5.40	5.44	5.44	5.46	5.63	5.63	5.67	5.64
	(0.06)	(0.04)	(0.07)	(0.04)	(0.05)	(0.03)	(0.05)	(0.03)	(0.06)	(0.03)	(0.07)	(0.03)
White girls	2.90	2.91	2.85	2.87	5.48	5.47	5.47	5.47	5.61	5.65	5.66	5.68
	(0.06)	(0.04)	0.06	0.04	0.05	0.03	0.05	0.03	0.06	0.03	0.06	0.03
Black girls	2.95	2.92	2.94	2.89	5.33	5.33	5.32	5.32	5.62	5.67	5.59	5.66
	0.06	0.04	0.06	0.04	0.06	0.03	0.06	0.04	0.06	0.03	0.07	0.03
Hispanic girls	2.97	2.96	2.91	2.88	5.34	5.36	5.33	5.35	5.69	5.63	5.66	5.65
	0.06	0.04	0.06	0.04	0.05	0.03	0.06	0.03	0.06	0.03	0.06	0.03
Black boys	3.02	3.02	2.99	2.99	5.49	5.43	5.46	5.40	5.68	5.68	5.73	5.71
	0.06	0.04	0.06	0.04	0.05	0.03	0.06	0.04	0.06	0.03	0.06	0.03
Hispanic boys	2.99	2.97	2.93	2.92	5.36	5.38	5.34	5.36	5.59	5.59	5.64	5.61
	0.06	0.04	0.06	0.04	0.05	0.03	0.06	0.04	0.06	0.03	0.07	0.04

Note: *N* = 450 teachers in the full sample and 390 teachers in the analytical sample. The mean scores reported here are based on models that controlled for each item, the positioning of the item on the questionnaire, and teacher fixed effects. Mean$_a$ is the mean score based on models that controlled for the correctness ratings together with the aforementioned variables. Mean scores can range from 1 (*very low mathematical ability*) to 7 (*very high mathematical ability*).