Stanford cepa | Center for Education Policy Analysis

CEPA Working Paper No. 18-03

Bias in Online Classes: Evidence from a Field Experiment

AUTHORS

Rachel Baker
University of California, Irvine

Thomas Dee
Stanford University

Brent Evans
Vanderbilt University

June John
Stanford University

ABSTRACT

While online learning environments are increasingly common, relatively little is known about issues of equity in these settings. We test for the presence of race and gender biases among postsecondary students and instructors in online classes by measuring student and instructor responses to discussion comments we posted in the discussion forums of 124 different online courses. Each comment was randomly assigned a student name connoting a specific race and gender. We find that instructors are 94% more likely to respond to forum posts by White male students. In contrast, we do not find general evidence of biases in student responses. However, we do find that comments placed by White females are more likely to receive a response from White female peers. We discuss the implications of our findings for our understanding of social identity dynamics in classrooms and the design of equitable online learning environments.

Acknowledgements: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B140009 to the Board of Trustees of the Leland Stanford Junior University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

VERSION

March 2018

Suggested citation: Baker, R., Dee, T., Evans, B., & John, J. (2018). Bias in Online Classes: Evidence from a Field Experiment (CEPA Working Paper No.18-03). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp18-03

Bias in Online Classes: Evidence from a Field Experiment

In educational settings, effective personal interactions (i.e., between instructors and students as well as among student peers) are potent mechanisms for driving student engagement and learning. However, there is evidence that the character and frequency of these interactions sometimes reflects biases that reinforce inequity. In particular, a long-standing literature finds that instructors in traditional educational settings interact differently with students based on the congruence of race, ethnicity, and gender (e.g., American Association of University Women, 1992; Farkas, 2003; Hall & Sandler, 1982; Rubovits & Maehr, 1973; Sadker, Sadker, & Klein, 1991). For example, in classroom interactions at the K-12 level, teachers, on average, direct more positive and neutral speech toward White students than toward Latinx and Black students, while directing similar amounts of negative speech at all students (Tenenbaum & Ruck, 2007). There is also experimental evidence that these biases exist even in settings that lack face-to-face interactions: college students with racially or gender-connotative names receive different responses from instructors when asking for a face-to-face meeting or when asking to discuss research opportunities as a prelude to applying for a doctoral program (Milkman, Akinola, & Chugh, 2012, 2015). These behavioral patterns could reflect implicit or unconscious biases (i.e., quick and reflexive judgments shaped by experience and culture but not conscious intent) as well as outright discriminatory attitudes.

Regardless of their cause, it is important to identify and mitigate race and gender biases, as they can meaningfully exacerbate educational inequality. In particular, the effects of biases on student achievement are suggested by an active and growing body of evidence linking the racial and gender congruence of instructors and students to student learning in K-12 and higher education (e.g., Dee, 2004, 2005, 2007; Fairlie, Hoffman, & Oreopoulos, 2014; Gershenson,

Holt, & Papageorge, 2016; Lindsay & Hart, 2017; Van den Bergh, Denessen, Hornstra, Voeten, & Holland, 2010).

Using a field-experiment, this study provides what we believe is the first evidence of the possible presence of racial and gender biases among students and instructors in online courses. Our experimental study is situated in the discussion forums of 124 Massive Open Online Courses (MOOCs). In online learning environments, such forums provide the primary, and often the only, opportunity for instructors and students to interact. These interactive message boards also perform vital educational functions as students rely on the discussion forums to ask questions about the course content and structure and to receive answers from fellow students and course instructors. We tested for the presence of racial and gender biases in these settings by creating fictional student identities with racial- and gender-connotative names, having these fictional students place randomly assigned comments in the discussion forums, and observing the engagement of other students and instructors with these comments.

Ex ante, it is not clear if these online settings will mitigate or increase biased interactions relative to in person educational settings. The comparative anonymity of these entirely digitally mediated interactions, which provide fewer visual clues of race or gender, could attenuate racial and gender biases by reducing the tendency towards the social categorizations that triggers biases. Alternatively, the online setting may increase biased interactions by reducing the social incentives for self-control.

To preview the results, we find that instructors (i.e., professors at selective universities) are 94% more likely to respond to a discussion forum post by a White male than by any other race-gender combination. In contrast, we find, in general, no biases in responses by students to comments placed by students with experimentally assigned identities. However, we do find

evidence that comments placed by a student with a particular race-gender identity are more likely to elicit a response from other demographically similar students (this is especially true for White female students).

We believe that there are at least three distinct contributions of this study. First, it provides novel and fundamentally important insights into a rapidly proliferating type of learning environment. In 2013, 25 percent of all postsecondary students took some or all of their courses online (McPherson & Bacow, 2015). This fact has equity implications given that students enrolling in less selective colleges make up a larger fraction of the online student body (McPherson & Bacow, 2015). Even in K-12 education, more than 300,000 students exclusively attend online schools, with as many as 5 million students having taken at least one online course (Samuelsohn, 2015). This trend is likely to continue as educational institutions simultaneously seek to expand access and to control costs. However, despite their rapid growth, we currently know relatively little about the challenges and opportunities for promoting equity in these digital learning spaces.

Second, our empirical evidence also makes distinct theoretical contributions. Much of the literature on biases in student-teacher interactions cannot cleanly separate instructor-centered effects (e.g., implicit biases) from student-centered phenomenon. Such student-centered reactions would include, for example, female and minority students experiencing poorer educational outcomes with a White male teacher, not because of biases in the teacher's behavior, but rather because the teacher's identity triggers educationally relevant reactions like stereotype threat (Steele & Aronson, 1995). Because our study relies on fictive student identities, it cleanly isolates behavioral effects due to instructors and unequivocally rules out mechanisms related to student reactions to a particular instructor. Additionally, the heterogeneity in our results (i.e., by

course and comment type, student, and instructor identity) provides indirect empirical evidence on the different theoretical perspectives that explain instructor bias.

Third, the focus of the quantitative literature on bias in education settings has been almost exclusively on instructor- student interactions and has ignored the potential biases between student peers. However, interactions with fellow students can meaningfully influence student outcomes, even in online settings (e.g., Bettinger, Liu, & Loeb, 2016), so patterns of bias in peer interactions are potentially important. In this study, we are able to observe the racial and gender identity for most of the students who responded to our experimentally designed identities and comments. These data allow us to examine whether a response to a student with a particular identity is more common for demographically similar students.

Our paper is organized as follows. We first discuss the empirical literature and theoretical perspectives on bias in education as well as its relevance for online education. We then describe our study context, design, data, and methods. After presenting and discussing our findings, we conclude with thoughts on their implications for our understanding of classroom equity in general as well as for the design of equitable online learning environments.[1]

Bias in Education

A large and growing body of evidence suggests that persistent biases in human judgment related to race and gender influence personal interactions in multiple domains of human activity such as health care (Saha, Kamaromy, Koespell, & Bindman, 1999), law enforcement (Gelman, Fagan, & Kiss, 2007), the housing market (Ahmed & Hammarstedt, 2008; Ewens, Tomlin, & Wang, 2014; Hanson, Hawley, Martin, & Liu, 2016), and the labor market (Bertrand &

[1] To be clear, we intend for equity to refer to the quality of being free from bias and favoritism but acknowledge that others instead interpret equity as differentiated inputs in support of equality of opportunity.

Mullainathan, 2004; Edelman, Luca, & Svirsky, 2017; Ewens, Tomlin, & Wang, 2014; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Oreopoulos, 2011). Racial minorities and/or women are consistently disadvantaged in each of these settings. Given these results, it is perhaps unsurprising that there is parallel empirical evidence involving race- and gender-based biases in every level of schooling.

In particular, a long-standing body of evidence indicates that students at all levels of education experience patterns of bias with respect to race, ethnicity, and gender in classrooms. For example, boys generally receive more attention and comments from instructors than girls in primary education (e.g., American Association of University Women, 1992; Sadker & Sadker, 1986). There is also evidence that teachers treat Black students more negatively than White students (Rubovits & Maehr, 1973) and reinforce social aspects of behavior for Black girls while highlighting academic behaviors of White girls (Damico & Scott, 1987). White teachers are also likely to rate Black students' misbehavior more harshly than similar behavior of White students (Downey & Pribesh, 2004). These problems are also documented internationally as racial biases exist in teachers' evaluations of ethnic-minority immigrants (Van den Bergh et al., 2010).

Interactions in postsecondary education are also subject to race and gender biases. Observational studies have noted that college faculty both overtly and subtly discriminate against women (Hall & Sandler, 1982), and McGee (2016) has documented racial micro aggressions of faculty against Black STEM students in postsecondary education. Experimental analyses have credibly identified the existence of faculty discrimination against women and racial minority applicants for lab positions and doctoral programs at the graduate level (Milkman, Akinola, & Chugh, 2015; Moss-Racusin et al., 2012). Although these biases are widespread, it remains to be seen if they exist between student and teacher interactions and between student and student

interactions in the virtual classroom where we might expect lower levels of bias due to the relative anonymity and lack of subsequent in-person interactions or higher levels if the anonymity in online environments reduces self-control.

Potential Causes of Bias in Education

In general, there are three broad theories of discrimination (National Research Council, 2004) that potentially explain why instructors and students might exhibit bias in classroom interactions.[2] One of the most prominent explanations - implicit or unconscious bias - reflects the claim that individuals carry (and sometimes act upon) the unconscious attribution of stereotypes to a particular social identity (e.g., Staats, Capatosto, Tenny, & Mamo, 2017; Dee & Gershenson, 2017). The literature on implicit bias is rooted in long-standing notions from the field of psychology that social cognition reflects, in part, automatic or unconscious processes (e.g., Shiffrin and Scheider, 1977; Devine, 1989) that are difficult to suppress voluntarily. This particular human tendency (i.e., to make quick, reflexive categorizations and decisions) is sometimes framed as an evolutionary adaptation (Kahneman, 2011). However, it is also understood that social and cultural forces (e.g., Rudman, 2004) can shape implicit social cognition in ways that instantiate discrimination (i.e., through cultivating involuntary and unconscious stereotypes).

A second category of theories involve "intentional, explicit discrimination" (National Research Council, 2004). At its most benign, the harm caused by such "taste-based discrimination" (Becker, 2010) begins with the avoidance of "outgroup" contact as well as verbal and nonverbal hostility. A third category involves statistical discrimination and profiling.

[2] The NRC report also identifies a fourth conceptual category (i.e., the ways in which discrimination can influence institutional processes and organizational rules) that has relevance in education but less so for our study of within-course behaviors.

Statistical discrimination refers to the notion that people discriminate against individuals because they consciously ascribe to that individual the average characteristics they attribute to the individual's social identity (Aigner & Cain, 1977; Arrow, 1998; Schwab, 1986). In the words of Phelps (1972, p.659), "skin color or sex is taken as a proxy for relevant data not sampled." For example, in the context of our study, an online instructor who statistically discriminates might be more likely to respond to a comment from a student with a particular social identity because they believe that identity is associated with higher achievement and that their question therefore signifies the likelihood of particularly widespread confusion in the class.

These conceptual frames are not only relevant for instructor behavior. They can also apply to patterns of bias in the peer-to-peer interactions among students. For example, students may exhibit preferences for engaging a student with a given social identity if statistical discrimination shapes the perceived value of such engagement. Additionally, network studies (e.g., McPherson, Smith-Lovin, & Cook, 2001) suggest that individuals consistently demonstrate preferences for engaging others who share their traits (i.e., "homophily"). These patterns may reflect the intergroup avoidance implied by intentional discrimination as well as implicit biases. In our study, we identify experimentally whether such biases exist in online courses. These experimental results do not directly test these different theories. However, after presenting our study design and main results, we discuss how the treatment heterogeneity in these findings provides indirect empirical evidence on these different mechanisms.

Consequences of Bias in Education

Although not all of the gender and racial gaps observed in education are caused by bias and discrimination, the differential interactions discussed above do appear to play a meaningful role in explaining differences in student outcomes (Mickelson, 2003). Hall and Sandler (1982)

argue that, on college campuses, differential treatment between men and women signals different expectations, which reduces the contribution of women in the class and dissuades women from studying certain fields. These problems may be particularly acute in STEM disciplines.

It is also widely held that differences in teacher and student race matches affect students' academic performance and contribute to the resulting racial achievement gaps (Dee, 2005; Ferguson, 2003; Van den Bergh, 2010). When students were randomly assigned a teacher whose race matched their own race, the achievement of both Black and White students improved (Dee, 2004). Racial matches also appear to affect teachers' academic perceptions of students (Dee, 2005) and teachers' social-emotional ratings of students (Wright, Gottfried, & Le, 2017). These racial match findings also extend into higher education. When minority students had a minority faculty member in community college, they experienced improved retention, academic achievement, and degree completion (Fairlie, Hoffman, & Oreopoulos, 2014). However, none of these studies can disentangle effects that are driven by student-centered behaviors and perceptions (e.g., role model effects) from effects that are driven by instructor behavior (e.g., due to implicit biases).

Student Engagement in Online Education

Discussions of equity in online education tend to focus on either how their comparatively low cost and online delivery can broaden access or, conversely, on how the uneven distribution of computer hardware and broadband connections inhibits the realization of this promise (i.e., the digital divide). However, we know of little evidence that examines issues of equity *within* online classrooms. Our study is motivated by the view that this is an important omission in the literature on online learning environments.

In conventional classrooms, the interactions among students and instructors are important determinants of student engagement, which in turn is a key mediator of educational success (Fredricks, Blumenfeld, & Paris, 2004). There is evidence to suggest that such interactions play an equally, if not more important, role in online settings (Bedlarrain, 2006; Dixson, 2010). In online classes, these interactions typically occur in discussion forums. Most online classes have an "asynchronous" design (i.e., students and their instructor do not interact simultaneously). Therefore, the discussion forums in such courses are the central environments in which students can engage with their instructor and each other (Hart, 2012). The evidence that an interactive community of inquiry is necessary to achieve success in online courses (Garrison & Cleveland-Innes, 2005) underscores the relevance of these forums. And the relevant interactions in these forums are not just between students and their instructor. A meta-analysis by Bernard et al. (2009) concluded that student-to-student interactions in distance and online courses are positively associated with various measures of student learning.

The relevance of discussion forums in online courses implies that these are the settings in which biased interactions with relevance for learning outcomes may or may not occur. This motivated our decision to situate our experimental study, which we describe below, in such forums. Understanding what types of bias may or may not exist in these settings is transparently relevant for concerns about equity in these new learning environments. Understanding the determinants of student engagement in online settings is also more generally relevant because these settings, especially Massive Open Online Courses (Evans, Baker, & Dee, 2016; Perna et al., 2014), often suffer from low in-course persistence.

Current Study

We study the presence and extent of student and instructor racial and gender biases in online discussion forums of online postsecondary courses. Specifically, we examine the presence of racial and gender bias in these online courses through the analysis of a large-scale randomized field experiment in which we posted comments in discussion forums using randomly assigned racial and gender-connotative names linked to student profiles we created. We observed both whether the instructor and the other students in these courses engaged with these designed comments. We situated our study within 124 Massive Open Online Courses. Despite the cycle of early hype and then cynicism around MOOCs, these free classes remain a widely used form of online learning. In 2017, more than 800 universities offered 9,400 unique MOOCs, and 78 million students signed up for at least one course (Shah, 2017). Importantly, MOOCs are playing a growing role in postsecondary credentialing; students can earn course credits or even certificates from accredited colleges through them. Critically, we also believe there is credible external validity to conducting this study within MOOCs because their basic design features (e.g., asynchronous engagement, recorded lectures, discussion forums) and their postsecondary content are widely used in other online courses.

Experimental Design

We identified our experimental sample of MOOCs by compiling the universe of MOOCs offered by a major provider that started between August 1 and December 31, 2014.[3] We screened the available courses and included those that met the following criteria: five weeks or longer, not targeted at children (i.e., under age 18), had a general discussion forum, and was not taught by an instructor that was included in our small preceding pilot. Additionally, we only included one

[3] As part of our human-subjects protocol, we do not identify this provider nor do we provide the titles of the classes or the exact text of the comments we placed.

course per instructor. When instructors taught more than one course, we decided which course to include based on date (i.e., taking earlier courses over later ones) and length (i.e., taking longer rather than shorter classes). When all else was equal, we took the course that was listed first alphabetically. The 124 MOOCs in our sample covered a diverse range of subjects, including accounting, calculus, epidemiology, teaching, and computer programming. Most (94) were offered by four-year not-for-profit institutions of higher education in the United States; those that were instead offered by international institutions were taught in English.

Using fictive student identities, we placed eight discussion-forum comments in each of the 124 MOOCs. Within each course, eight student accounts were used to place one comment each. The eight student accounts each had a name that was connotative of a specific race and gender (i.e., White, Black, Indian, Chinese, each by gender); each race-gender combination was used once per class. Our random-assignment procedure, which we describe below, was designed to ensure that the student name, the comment they placed, and the order in which each comment was placed was placed were random. We placed comments in the "General Discussion" or similar sub-forum and we timed comments to be spaced out roughly equally over the duration of the course, from the beginning of the course to two weeks before the end of the course. We observed all replies to each comment for the two weeks after placement.[4] By observing the responses to our comments by instructors and by students in the course, we can identify any difference in the number of responses received by our student accounts that were assigned different race and gender identities.

Comments. Drawing from several hundred actual student comments placed in a variety of MOOCs, we constructed a list of 32 generic discussion forum comments that would be

[4] Our small pilot study that preceded the experiment indicated that this window would capture the responses to virtually all placed comments.

applicable across all types of courses. Our comments focused on topics such as praise for the course or instructor, questions about studying, and issues of course difficulty that could be sensibly placed in any course regardless of the subject matter. Some of the comments focused on issues directly related to course procedures (e.g., specific questions about due dates or questions about how to complete assignments), and we phrased these specific comments such that it appeared the comment poster needed an answer in order to be able to successfully complete the course. We refer to this set of comments as "completion-focused." Other comments were declarative statements that might catalyze conversation (e.g., a comment that the course was easier than the student expected) or questions about other students in the class (e.g., asking where people are from or why they are taking the class). In this second group, a poster's course success did not hinge directly on getting a reply to the comment. We refer to this set of comments as "advising/social." A description of all 32 comments and their categorization can be found in Appendix A.[5] On average, the frequency of student and instructor response to our comments was similar to that of real student comments in our MOOCs, suggesting that our comments were representative and realistic.

Names. We randomly paired comments to students with our race-gender evocative names. To create a bank of names, we drew from Anglo-American, African-American, Indian, and Chinese names that were recently used in studies that have also experimentally manipulated perceptions of race and gender (Bertrand & Mullainathan, 2004; Milkman, Akinola, & Chugh, 2015; Oreopoulos, 2011). We identified a set of four first names and four last names for each gender in each race (16 unique names for each of 8 race-gender combinations, 128 unique names in total). Each posting used a first and last name, which is a common practice by actual students

[5] In order to preserve the anonymity of those engaged in this experiment, we edited these comments slightly (but without changing their meaning) so they could not be identified. Appendix Table A1 also identifies alternative comment classifications we use as a robustness check.

in MOOC forum postings, to maximize the chance of being identified with the appropriate race-gender profile.

Randomization. In each MOOC, we had one of each of our eight race-gender identities place one randomly assigned comment. This within-course design allows us to control unrestrictedly for all the unobserved course-specific traits that may influence commenting within the course. However, to avoid other potential confounds, we also adopted procedures that would create random variation in both the comment placed (i.e., which of the 32 comments) and the order in which it was placed in the course (i.e., 1st through 8th). For example, to choose the sequencing of race-gender profiles within each course, we first established an initial random ordering of the sequence of the eight race-gender profiles and did so in a manner that ensured that no same-gender or same-race identity appeared consecutively. For the first course in our study, we then randomly assigned 8 comments to these profiles in this randomly ordered sequence (i.e., 1, 2, 3,…, 8). We also randomly assigned one of the 16 possible names appropriate for the race-gender identity of each poster.

These 8 initial comments were randomly selected without replacement from the total list of 32 comments. When a second eligible course opened, we randomly selected 8 comments from the remaining pool and assigned them to race-gender profiles in a sequence that was rotated by one position (i.e., 2, 3,...,8, 1). As subsequent courses opened, we randomly selected matched comments until the pool of 32 was exhausted. After every four courses, our procedure returned to the full set of 32 comments. Similarly, we continued rotating the sequence in which race-gender profiles appeared and re-randomized when a full rotation was achieved (i.e., every 8 courses). We also relied on random selection of names without replacement and then re-randomized every 16 times so that names were balanced in the design of the study.

This process has several important features. First, it guarantees, for all participating courses, within-course variation in the student identities placing comments (i.e., the "treatment" of interest in our experiment). Second, by design, it also provides random variation for each student identity posting *within* courses in both the comment placed and the order in which it was placed. Finally, our approach ensures across all the courses a balanced representation of all the identities, names, and comments used in our study. We observe this balance in our final data set. For example, **e**ach particular race-gender profile (i.e. White male) was used exactly once per course, so each was used 124 times. The number of times a particular name in each race-gender profile was used ranged from 6 to 8. The number of times each of the 32 comments was used across the entire study ranged from 29 to 32 with each race-gender profile placing each comment an average of 3.9 times.[6]

In a conventional experimental study, an important check on the study design is to examine whether the observed traits of the participating subjects are well balanced across treatment and control conditions. The issue of covariate balance has less relevance in our study because our observations (i.e., the placed comments in these online classes) have no covariates beyond our randomly assigned treatments of interest (i.e., the race/gender identity) and the categorical traits (i.e., course, comment type, comment sequence) for which we provide unrestrictive controls through the use of fixed effects. Nonetheless, to provide further evidence on the covariate balance in our design, we regressed a dummy variable representing each race and gender identity on courses fixed effects and a set of fixed effects for comment type and the comment order (i.e., sequence). For 7 out of these 8 auxiliary regressions (Appendix Table B1),

[6] The slight imbalance in the frequency of names used and comments relative to what our design would imply is due to the fact that we dropped two courses in which we had begun placing comments. One course was dropped because our monitoring of student comments raised concerns that the existence of our study might be uncovered. A second course was dropped because, unlike other courses, it ceased accepting new registrants during the course progression. Including the data that we did collect from these courses does not influence our findings.

F-tests indicate that we cannot reject the null hypothesis that these comment and sequence fixed effects have no "effect" on the assigned racial and gender identity of the poster.[7]

Analysis

Our analytical strategy closely parallels our experimental design. That is, we regress our key outcomes (i.e., measures of student and instructor responses) on seven race-gender indicators (i.e., using White male as the reference category) and conditional course, comment, and sequence fixed effects. Our preferred model is:

$$Y_{ijkt} = \alpha + \sum_{i=2}^{8} \beta_i R_i + \theta_j + \delta_k + \mu_t + \epsilon_{ijkt}$$

where Y_{ijkt} is the outcome for posting *i* of comment *k* placed in the discussion forum of course *j* in the t^{th} position of the sequence of comments in that course. R_i refers to the assigned race-gender profile of the comment. The term, θ_j, is a course fixed effect. The term, δ_k, is a comment fixed effect, and μ_t is a sequence fixed effect for the order in which the comment appeared. We allow the error term, ε_{ijkt}, to reflect the nestedness of the comments within courses by clustering the resulting standard errors at the course level.

These course, comment, and sequence fixed effects account unrestrictedly for the natural heterogeneity in outcomes by the course, sequence order of the comment, and text of the comment. That is, they control for all variables that are constant within a course (e.g., general frequency of discussion forum activity), the average number of responses each particular comment receives across all courses, and the average effects of placing a comment earlier or

[7] The one exception is for female Chinese identities. A closer inspection revealed that this spurious correlation is due to our randomization causing the female Chinese identities to be linked to some comments as few as 0 times and other comments as often as 8 times. However, it should be noted that we condition on comment fixed effects; also, we observe qualitatively similar findings when we drop all female Chinese observations.

later in a course. While the randomization we describe above should control for any concerns about differences in response rates across courses, comments, or the timing of comments, these fixed effects further ameliorate any potential poor randomization. For example, if the Black female profiles are randomly assigned to place the first comment (which is more likely to receive a response) in classes with very active discussion forums more often than other race-gender combinations, these fixed effects will control for effects related to being in an active course as well as effects related to placing the first comment.

There are three main outcomes in our confirmatory analysis: whether an instructor replied to the comment, whether at least one student replied to the comment, and the total number of students who replied to a comment.[8] As an exploratory exercise, we also estimate the same model above on different subgroups of courses and comments to explore the treatment heterogeneity in our study. For ease of interpretation, we use a linear probability model to estimate the effect of race-gender profile on the likelihood of response. Estimated effects from logit models for binary outcomes and negative binomial models for count outcomes produce similar results. These results are available upon request.

Results

We have a total of 992 postings (8 individual comments placed across 124 courses), each of which was assigned one of eight treatments (race-gender). We received a total of 3,588 replies, made by 2,976 unique users. Descriptive statistics for responses to our comments are provided in Table 1. Instructors replied to 7.0% of our comments. At least one student responded

[8] This design implies that our main confirmatory evaluation involves estimating 21 point estimates (i.e., 7 race-gender identities across three outcomes). In Appendix C, we present evidence on whether our results may suffer from a "multiple comparisons" problem. Specifically, we reconsider the significance of our findings after adjusting for a "false discovery rate" (Benjamini and Hochberg, 1995). We also note that we consider our other inferences (i.e., treatment heterogeneity and student homophily) as exploratory (Schochet, 2008).

to 69.8% of our comments with an average of 3.2 student replies to each of our comments. The variance in the number of student replies to each comment is large with comments garnering between zero and 213 replies.

The remainder of Table 1 provides descriptive characteristics of the courses and comments in the study. STEM courses comprise 56.5% of the 124 courses in the sample. Fifty eight percent of the courses in our sample were taught by either one White male instructor or a teaching team of exclusively White men. We consider 43.6% of the comments to be focused on course completion with the remainder categorized as general advising or social comments. The poster identity rows demonstrate that we had balance across each race-gender combination; each race-gender profile posted exactly one comment in each course.

Presence of Instructor and Student Bias

Our analyses focus on two binary measures, whether an instructor replied and whether a student replied, and one continuous measure, the number of student replies, to each of our posted comments.[9] As a first step in the data analysis, we examine instructor and student response rates for our race-gender groups visually. Figure 1 presents unconditional probabilities of instructor response for each of our race-gender profiles. We observe that comments posted by White males appear more likely than all other student groups to receive a response from course instructors. As noted in Table 1, the overall rate of instructor response is 7% (indicated by the horizontal red line in the figure); however, over 12% of our comments posted by White men elicited a response from an instructor, and the rate of response is far lower for every other race-gender combination.

There are several ways to assess whether the observed difference is statistically significant. One simple way is to test whether the observed distribution in Figure 1 is different

[9] We exclude a continuous measure of number of instructor responses given that the median number is 0.

from a uniform distribution in which all bars are the same height. This chi-squared test fails to reject the null hypothesis that the data are drawn from a uniform distribution ($\chi^2(7) = 8.56, p = 0.285$). However, a simple t-test comparing instructor response rates to comments by White males versus the other student identities combined rejects the null hypothesis that these response rates are the same ($|t| = 2.41$, $p = 0.008$). Figure 2 demonstrates high and consistent response rates from students, as opposed to instructors, across race-gender categories. Again, a chi-squared test fails to reject the null hypothesis that these data are from a uniform distribution ($\chi^2(7) = 1.36, p = 0.987$). Overall, these unconditional probabilities suggest the existence of a race-gender bias among instructors (i.e., favoring White male identities) but do not suggest bias by other students in online educational discussion forums. One limitation of these simple tests is that they account neither for the blocked nature of the random assignment nor for other important controls that improve precision in the regression analysis below.

We formalize these descriptive findings using the regression model described above which controls for course, comment, and sequence fixed effects. Results are presented in Table 2 for our three different outcomes. For each outcome, we provide results from two regression models. The first model includes indicator variables for each race-gender combination using White males as the omitted category. In this specification, we also report the results of an F-test of the null hypothesis that the seven coefficients of interest are equal. The second model uses only an indicator for W males, effectively collapsing all the other race-gender combinations into the reference category.

The first model for the binary instructor reply outcome shows that the coefficients on *all* race-gender groups are quite large and negative (as compared to White males, the omitted category) with four of the seven comparisons being statistically significant. An F-test of the

hypothesis that these seven coefficients are the same is not rejected ($p = 0.477$), which gives us confidence that we can group these race-gender categories together. When we collapse the race-gender profiles into a comparison of White males versus all others, we see that a comment from a White male is a statistically significant 5.8 percentage points more likely to receive a response from an instructor than non-White male students. The magnitude of this effect is striking. Given the instructor reply rate of 6.2 percent for non-White male posters, the White male effect represents an 94 percent increase in the likelihood of instructor response.

Moving to the binary student reply outcome, we observe no consistent pattern of positive or negative coefficients, and only the White female category is statistically significantly different from White males. Comments by fictive students assigned White female names experienced a 12.9 percentage point increase the likelihood of receiving at least one response from a student. However, pooling the non-White men comments and comparing them to White men shows no statistically significant difference. When examining the number of student replies as a continuous outcome in the final column of Table 2, we observe no consistent pattern and no statistically significant results. White men did not receive more responses from students than any of our other race-gender profiles.

In sum, our results show compelling experimental evidence that instructor discrimination exists in discussion forums of online classrooms. Simply attaching a name that connotes a specific race and gender to a discussion forum post changes the likelihood that an instructor will respond to that post. Comments posted by White males garner more frequent instructor response than comments posted by any other race-gender profile in our study. The magnitude of this result is both statistically significant and practically important: the response rate is nearly *twice* as large for White males as it is for other race-gender groups. There is little evidence of differential

response among students except that White females are more likely than White men to receive a response from a student peer.

Treatment Heterogeneity

In Table 3, we report exploratory evidence on how the key findings from Table 2 vary by several instructor, course, and comment traits. Specifically, we examined the effect of a White male identity (i.e., relative to the other 7) on each of the three outcomes in samples defined by particular traits such as whether the instructor was a White male, whether the course was on a STEM topic, and whether the comment was focused on course completion or on more general advising or social topics.[10] To ease comparisons, we replicate the model 2 findings for our three outcomes for the full sample in the first row of Table 3. For the two outcome variables reflecting student engagement with the comments (i.e., the results in columns 2 and 3 of Table 3), we consistently find no evidence of statistically significant effects across these different subsamples.

However, with regard to the probability that the instructor responded to the comment, we find several interesting patterns. For example, we find that the effect of a White male identity on the probability the instructor responded is larger when the instructor is also a White male. In courses that are not taught by a White male, the effect of a White male identity is smaller and statistically insignificant. We find no appreciable difference in the effect of a White male identity across STEM and non-STEM courses. However, we do find that this effect is larger among comments that are focused on advising or social questions and statistically insignificant with regard to comments that are focused on course completions.

[10] Five of our 124 courses are taught by multiple instructor teams of mixed race. We consider those courses to be non-White male instructor courses.

We caution against overintepreting these heterogeneity findings as these differences are not themselves statistically significant.[11] Nonetheless, these patterns do provide some weakly suggestive evidence on the theoretical mechanisms that may characterize our findings. For example, the results in both Tables 2 and 3 argue against statistical discrimination as an explanation for the instructor bias in favor of White males. There is evidence from educational research that instructors view certain groups of students, particularly male, White, and Asian students, as more able and higher achieving than other groups of students (Ferguson, 2003; Hsin & Xie, 2014; Kao, 1995; Riegle-Crumb & Humphries, 2012; Tiedemann, 2002; Wong, 1980).[12] However, if instructors were exhibiting behavioral biases because of statistical discrimination, we might expect a different pattern of heterogeneity across the 7 identities. For example, in Table 2, the estimated effects of a Chinese male and Black male identity are virtually identical. If instructors were statistically discriminating, the relevant stereotypes are unlikely to produce such a result.

Furthermore, one could conjecture that statistical discrimination by instructors is more likely to occur in STEM for at least two reasons. First, women and racial minorities are generally underrepresented in STEM fields. Second, there is evidence that instructors believe Asian and male students are likely to be higher performing in math than other groups (Cherng, forthcoming; Riegle-Crumb & Humphries, 2010). The fact that the effect of a White male identity is similar across STEM and non-STEM courses (Table 3) also argues against statistical discrimination as an explanation for the biased instructor behavior we observe.

[11] We examine these forms of heterogeneity in models that are based on the full sample and that allow interactions between the White male identity and these instructor, course, and comment traits.

[12] It is important to point out that research suggests that these beliefs do not match the empirical truth. In the limited demographic evidence of MOOC students, women perform just as well as men (DeBoer, Stump, Seaton, & Breslow, 2013), and there is no evidence on differential performance by race.

The other results in Table 3 weakly suggest implicit bias among these online instructors as the relevant mediator. For example, consider the fact that a White male identity leads to particularly large increases in instructor responses when the comment is advising or social rather than focused on course completion.[13] Comments that are narrowly focused on the course may catalyze more deliberate (i.e., bias free) responses from instructors because they are a core instructional responsibility. In contrast, with comments that are advising or social in nature (e.g., "Where does everybody come from?"), instructors are likely to feel that the decision to respond is more discretionary. The instructors, who are predominantly White males, may be more likely to respond to these comments when placed by White males because they are unconsciously more comfortable with such "ingroup" contact.[14] Furthermore, the evidence that the prevalence of bias varied across these two categories of comments argues somewhat against explicit or intentional discrimination.

Student Homophily

Our experimental data also enable us to explore student homophily. Although we find little evidence of real online MOOC students differentially replying, on average, to comments posted by students of different race and genders, that result does not preclude the possibility of students preferring to respond to comments posted by people who share their race and gender. By observing the public online profiles and names of the real students who responded to our comments, we can test whether gender and/or racial homophily exists among students in online

[13] We acknowledge that our classification is not the only way to divide our comments. To ensure the robustness of these results, we examined several alterative categorizations of these course comments (also noted in Appendix Table A1). All groupings resulted in qualitatively similar results.

[14] Indeed, when we limit the sample to the majority of courses that are taught by White males, we find an even larger effect of a White male identity with respect to the instructor responding to a "social/advising" type of comment.

educational discussion forums.[15] We accomplish this test by constructing a series of outcome variables that measure whether the comment posting received a reply (and, for a second outcome, the number of replies) from peers of matching race and/or gender. We then regress each of these outcomes on receiving a response (or number of responses) from students of that specific race and/or gender and include comment, sequence, and course fixed effects. This model is similar to our main estimating equation except we run a separate equation for each gender and/or race. Table 4 contains coefficients and point estimates from each of these regressions.

As an example of interpreting coefficients from this Table, the 0.059 estimated coefficient in the first row, first column implies that White students were 5.9 percentage points more likely than non-White students to respond to one of our comments when that comment was assigned a White name. We observe several marginally significant results throughout the table indicating the presence of homophily among female, White, and Indian subgroups. However, the only large and highly statistically significant result is among White female students responding to White female posters. We find that White women were over 10 percentage points more likely to respond to a post with a White female name than non-White women.

Discussion & Conclusion

In this study, we report novel field-experimental evidence that the equity concerns that are widely discussed in regard to conventional classrooms also exist in online learning environments. In other words, we find that online learning environments are still social

[15] We determined real student race and gender in three ways. First, we observed the public profiles of respondents to our comments. If a race and gender were provided in that public profile, we rely on the stated race and gender. Second, if the public profile did not state a race and gender but provided a picture, we use the picture to determine race and gender. Third, if the absence of other information, we use student first and last names, which are commonly affiliated with discussion forum postings, to guess the student's race and gender. Members of our research team coded the race and gender of each name using their best judgment and publically available lists of names. Our research team agreed on the gender and race of 64% of repliers to our comments.

environments in which identities can have salience. We situated our field experiment in the discussion forums of online courses. Because online courses are typically asynchronous, these forums provide a uniquely important venue for instructor-to-student and student-to-student engagement. Our field experiment produced evidence that the comparative anonymity granted by asynchronous, digitally mediated interactions in online discussion forums does not eliminate bias among instructors. Indeed, we found a sizable bias in favor of White male identities which were nearly twice as likely to receive a discussion-forum response from the instructor compared to other student identities. Furthermore, while we found no corresponding evidence of a general bias in peer-to-peer interactions among students, we did find evidence of homophily among some student groups (i.e., particularly White females).

We believe our findings also make an important contribution to the broader and quite active literature on the effects of race and gender-congruent instructors. These studies generally suffer from a limitation that attenuates their specific guidance for policy and practice. That is, these studies cannot cleanly identify the extent to which the effects of a "teacher like me" are due to student-centered effects (e.g., role model effects, stereotype threat) and/or instructor-centered effects (e.g., bias). Because our study relies on experimentally constructed student identities, it unambiguously isolates the effects that are instructor-centered. Furthermore, we are also able to discuss how the heterogeneity in our findings is most consistent with the specific hypothesis that these instructor behaviors reflect implicit bias (i.e., rather than intentional bias or statistical discrimination). While this evidence does not preclude the relevance of student-centered effects, it does suggest that teacher-facing interventions that reduce biased behaviors are likely to be both well targeted and effective in supporting student engagement.

Despite the advantages of our field-experimental approach, at least two caveats are notable. First, we intentionally chose names based on their clear affiliation with a race-gender profile. Students with names less easily associated with a specific race-gender may face less discrimination. Second, because our forum posters are fictive, we cannot assess the effects that the biases we observe may have on student performance or persistence in the course. Because the instructor and peer-engagement measures we study are in all likelihood important mediators of learning outcomes, we suspect that such effects exist. However, examining the effects of bias on student outcomes in online settings will require further and different study.

For example, one broad and possibly compelling direction would be to design, implement, and evaluate alternatively designed online learning environments that are effective in promoting equitable forms of engagement. Relative to conventional classrooms, online environments are uniquely amenable to such design innovations, in part because they can be implemented at scale with both fidelity and relatively little cost. For example, one obvious and simple approach would be to structure these classrooms in a manner that kept student identities strictly anonymous (e.g., removing names and photos). However, we also note that such extreme anonymity may have unintended consequences. A more sophisticated approach would be to structure online environments that guide instructors to engage with students in more equitable ways (e.g., dashboards that provide real-time feedback on the characteristics of their course engagement or short, embedded professional-development modules). The design features of online learning environments can also be adapted to either reduce homophily among students or to promote it when it aligns with educational goals. Regardless, our field-experimental study suggests such design innovations merit careful consideration given the evidence of biases our study uncovered.

References

Ahmed, A. M., & Hammarstedt, M. (2008). Discrimination in the rental housing market: A field experiment on the internet. *Journal of Urban Economics*, 64 (2), 362-372.

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

American Association of University Women (AAUW). (1992). *How Schools Shortchange Girls.* Washington, D.C.: AAUW Educational Foundation.

Arrow, K. J. (1998). What has economics to say about racial discrimination? *Journal of Economic Perspectives*, 12, 91-100.

Becker, G. S. (2010). *The economics of discrimination*. Chicago, IL: University of Chicago Press.

Beldarrain, Y. (2006). Distance education trends: Integrating new technologies to foster student interaction and collaboration. *Distance Education*, 27, 139-153.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bernard, R., M., Abrami, P. C., Borokhobski, E., Wade, C. A., Tamim, R. M., Surkes, M. A., & Bethel, E. C. (2009). A meta-analysis of three types of interaction treatments in distance education. *Review of Educational Research*, 79, 1243-1289.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review,* 94, 991-1013.

Bettinger, E., Liu, J., & Loeb, S. (2016). Connections matter: How interactive peers affect students in online college courses. *Journal of Policy Analysis and Management*, 35, 932-954.

Cherng, H. S. (forthcoming). If they think I can: Teacher bias and youth of color expectations and achievement. *Social Science Research*.

Damico, S. B. & Scott, E. (1987). Behavior differences between black and white females in desegregated schools. *Equity & Excellence in Education*, 23, 63-66.

DeBoer, J., Stump, G.S., Seaton, D., & Breslow, L. (2013). Diversity in MOOC students' background and behaviors in relationship to performance in 6.002x. *Proceedings of the Sixth Learning International Networks Consortium Conference*.

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

Dee, T.S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95, 158–165.

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

Dee, T.S., & Gershenson, S. (2017). Unconscious Bias in the Classroom: Evidence and Opportunities. Mountain View, CA: Google Inc. Retrieved from https://goo.gl/O6Btqi.

Devine, P., 1989, "Stereotypes and prejudice: Their automatic and controlled components", *Journal of Personality and Social Psychology*, 56: 5–18.

Dixson, M. D. (2010). Creating effective student engagement in online courses: What do students find engaging? *Journal of the Scholarship of Teaching and Learning*, 10, 1-13.

Downey, D. B., & Pribesh, S. (2004). When race matters: Teachers' evaluations of students' classroom behavior. *Sociology of Education*, *77*(4), 267-282.

Edelman, B., Luca, M. & Svirsky, D. (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*, 9, 1-22.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* (Vol. 1). Cambridge University Press.

Evans, B. J., Baker, R. B., & Dee, T. S. (2016). Persistence patterns in Massive Open Online Courses (MOOCs). *Journal of Higher Education*, 87, 206-242.

Ewens, M., Tomlin, B. & Wang, L.C. (2014). Statistical Discrimination or Prejudice? A Large Sample Field Experiment," *The Review of Economics and Statistics* 96 (2014), 119-134.

Fairlie, R. W., Hoffman, F., & Oreopoulos, P. (2014). A community college instructor like me: Race and ethnicity interactions in the classroom. *American Economic Review*, 104, 2567-2591.

Farkas, G. (2003). Racial disparities and discrimination in education: What do we know, how do we know it, and what do we need to know? *Teachers College Record*, 105, 1119-1146.

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the black-white test score gap. *Urban Education*, 38, 460-507.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59-109.

Garrison, R. D. & Cleveland-Innes, M. (2005). Facilitating cognitive presence in online learning: Interaction is not enough. *American Journal of Distance Education*, 19, 133-148.

Gelman, A., Fagan, J. & Kiss, A. (2007). An analysis of the New York City Police Department's "Stop and Frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102, 813-823.

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, *52*, 209-224.

Hall, R. M. & Sandler, B.R. (1982). *The classroom climate: A chilly one for woman?* Washington D.C.: Association of American Colleges.

Hanson, A., Hawley, Z., Martin, H., & Liu, B. (2016). Discrimination in mortgage lending: Evidence from a correspondence experiment. *Journal of Urban Economics*, 92, 48-65.

Hart, C. (2012). Factors associated with student persistence in an online program of study. *Journal of Interactive Online Learning*, 11, 19-42.

Hsin, A. & Xie, Y. (2014). Explaining Asian Americans' academic advantage over whites. *Proceedings of the National Academy of Science*, 111: 8416-8421.

Kao, G. (1995). Asian Americans as model minorities? A look at their academic performance. *American Journal of Education*, 10: 121-159.

Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux: New York, New York.

Lindsay, C. A., & Hart, C. M. (2017). Exposure to same-race teachers and student disciplinary outcomes for Black students in North Carolina. *Educational Evaluation and Policy Analysis*, *39*(3), 485-510.

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.

McGee, E. O. (2016). Devalued Black and Latino racial identities: A by-product of STEM college culture? *American Educational Research Journal*, 53: 1626-1662.

McPherson, M. S., & Bacow, L. S. (2015). Online higher education: Beyond the hype cycle. *Journal of Economic Perspectives*, 29, 135-153.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27: 415-444.

Mickelson, R. A. (2003). When are the racial disparities in education the result of racial discrimination? A social science perspective. *Teachers College Record*, 105, 1052-1086.

Milkman, K. L., Akinola, M., & Chugh, D. (2012).Temporal distance and discrimination: An audit study in academia. *Psychological Science*, 23, 710-717.

Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology*, 100, 1678-1712.

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

National Research Council. (2004). *Measuring Racial Discrimination*. Panel on Methods for Assessing Discrimination. Rebecca M. Blank, Marilyn Dabady, and Constance F. Citro, Editors. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Oreopoulos, P. (2011). Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy*, 3, 148-171.

Perna, L. W., Ruby, A., Boruch, R. F., Wang, N., Scull, J., Ahmad, S., and Evans, C. (2014). Moving through MOOCs: Understanding the progression of users in Massive Open Online Courses. Educational Researcher, 43, 421–432.

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

Riegle-Crumb, C. & Humphries, M. (2012). Exploring bias in math teachers' perceptions of students' ability by gender and race/ethnicity. *Gender & Society*, 26: 290-322.

Rubovits, P. C., & Maehr, M. L. (1973). "Pygmalion black and white," *Journal of Personality and Social Psychology*, 25, 210–218.

Rudman, L. A. (2004). Sources of implicit attitudes. *Current Directions in Psychological Science*, *13*(2), 79-82.

Sadker, M., & Sadker, D. (1986). Sexism in the classroom: From grade school to graduate school. *Phi Delta Kappan*, 67, 512-515.

Sadker, M., Sadker, D. & Klein, S. (1991). The issue of gender in elementary and secondary education. *Review of Research in Education*, 17, 269-334.

Saha, S., Komaromy, M., Koepsell, T. D. & Bindman, A. B. (1999). Patient-physician racial concordance and the perceived quality and use of healthcare. Archives *of Internal Medicine*, 159, 997-1004.

Samuelsohn, D. (2015, September 23). Virtual schools are booming. Who's paying attending? *Politico*. Retrieved from http://www.politico.com/agenda/story/2015/09/virtual-schools-education-000227.

Schwab, S. (1986). Is statistical discrimination efficient? *American Economic Review*, 76, 228-234.

Schochet, Peter Z. (2008). *Technical Methods Report*: *Guidelines for Multiple Testing in Impact Evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shah, D. (2017). By The Numbers: MOOCS in 2017 - Class Central. Retrieved from https://www.class-central.com/report/mooc-stats-2017/

Shiffrin, R. & W. Schneider, 1977, "Controlled and automatic human information processing: Perceptual learning, automatic attending, and a general theory", Psychological Review, 84: 127–190.

Staats, C., Capatosto, K., Tenny, L. & Mamo S. (2017). State of the science: Implicit bias review 2015 (Vol. 5). Kirwan Institute for the Study of Race and Ethnicity, The Ohio State University.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology*, *69*(5), 797.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology, 99*(2), 253-273.

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics*, 50: 49-62.

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.

Wright, A., Gottfried, M. A., & Le, V. (2017). A kindergarten teacher like me: The role of student-teacher race in social-emotional development. *American Educational Research Journal*, 54 (1S), 78S-101S.

Wong, M. G. (1980). Model students? Teachers' perceptions and expectations of their Asian and white students. *Sociology of Education*, 53: 236-246.

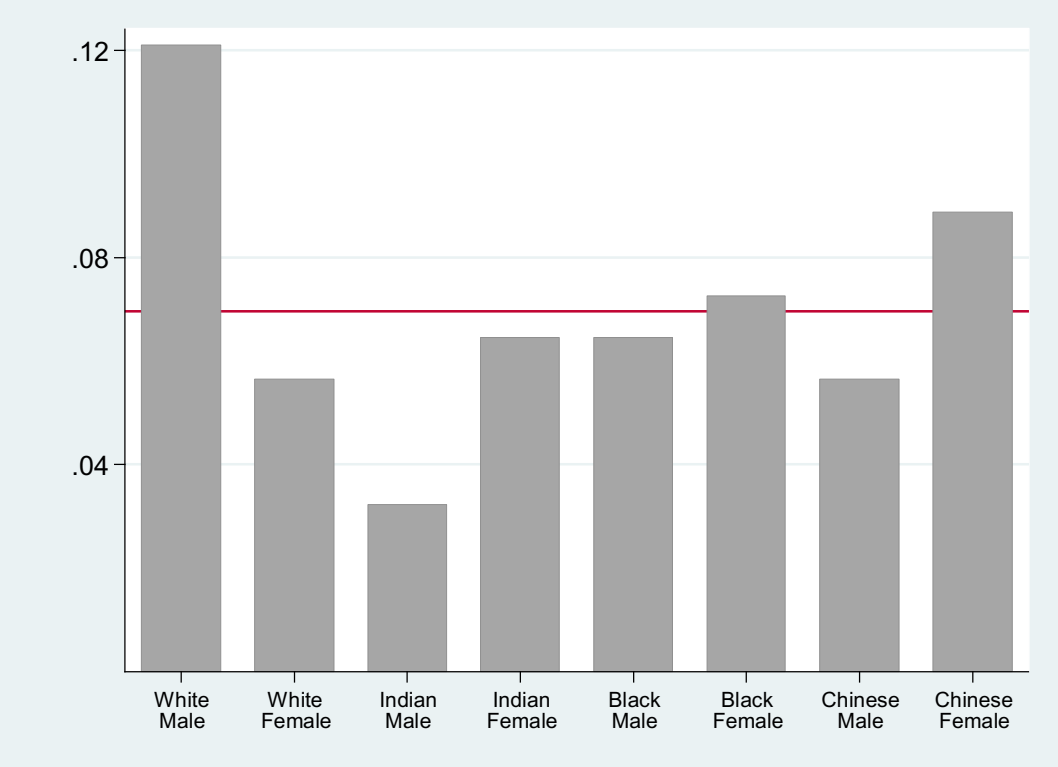

Figure 1 - Unconditional Probability of an Instructor Response by Student Identity

Notes: A chi-squared test cannot reject the hypothesis that the data are from a uniform distribution ($\chi^2(7) = 8.56$, p =0.285). A t-test rejects the hypothesis that instructors respond to White male students at the same rate as all other students ($|t|$= 2.41, p = 0.008).

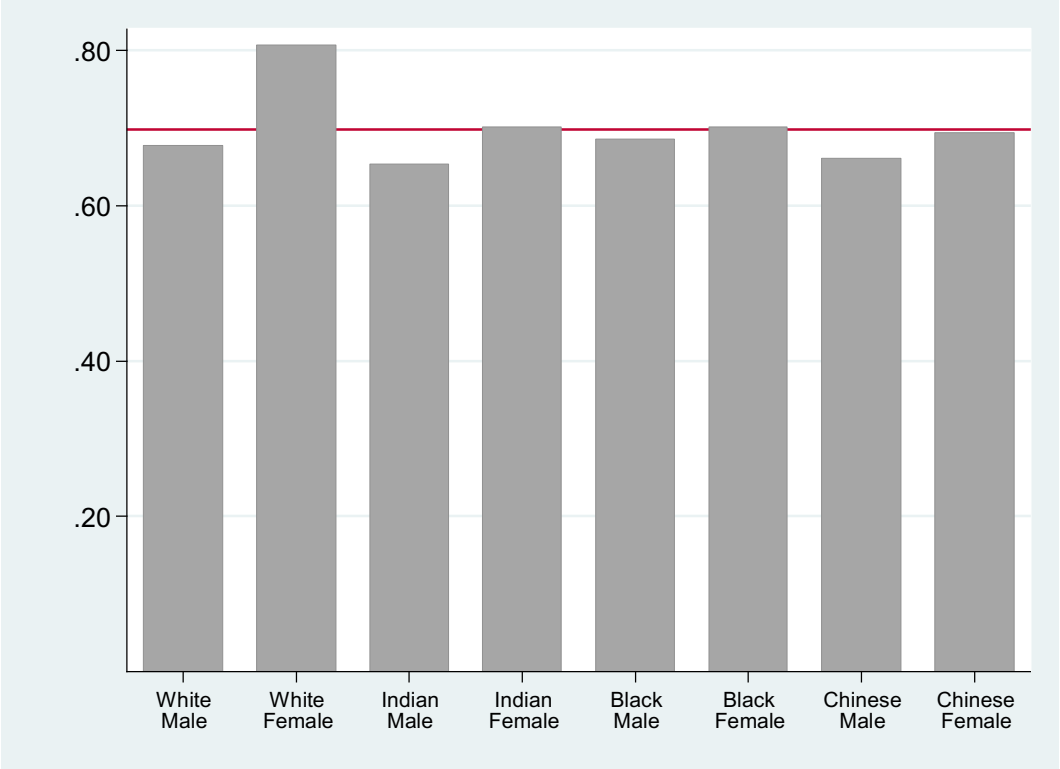

Figure 2 - Unconditional Probability of a Peer Response by Student Identity

Notes: A chi-squared test cannot reject the hypothesis that the data are from a uniform distribution ($\chi^2(7) = 2.79$, p =0.903).

Table 1 - Descriptive Statistics

Variables	Mean	SD	Min	Max
Outcomes				
Instructor Replied (0/1)	0.070	0.255	0	1
Student Replied (0/1)	0.698	0.460	0	1
Number of Student Replies	3.205	9.817	0	213
Course/Comment Characteristics				
STEM Course	0.565	0.496	0	1
White-Male Instructor	0.581	0.494	0	1
Completion-Focused Comment	0.436	0.496	0	1
Poster Identity				
White Male	0.125	0.331	0	1
White Female	0.125	0.331	0	1
Black Male	0.125	0.331	0	1
Black Female	0.125	0.331	0	1
Indian Male	0.125	0.331	0	1
Indian Female	0.125	0.331	0	1
Chinese Male	0.125	0.331	0	1
Chinese Female	0.125	0.331	0	1

Notes: The unit of observation is a comment placed in the discussion forums of online courses (i.e., 8 comments in each of 124 courses, N=992). The poster identity, the comment placed, and their sequencing were randomly assigned. See text for details. White-male instructor courses include single instructor courses taught by a white male and multiple instructor courses taught exclusively by white males. Non-completion-focused comments are comments labeled advising/social. See Appendix Table A1 for comment categorization.

Table 2 - The Estimated Effects of Student Identities on Instructor and Peer Responses

	Dependent Variables					
Independent Variable	Instructor Replied		Student Replied		Number of Student Replies	
White Male	-	0.058*	-	-0.021	-	-0.590
		(0.025)		(0.041)		(0.495)
White Female	-0.069*	-	0.129*	-	1.391	-
	(0.034)		(0.053)		(0.937)	
Black Male	-0.055+	-	0.011	-	-0.281	-
	(0.029)		(0.054)		(0.614)	
Black Female	-0.046	-	0.035	-	0.158	-
	(0.032)		(0.055)		(0.638)	
Indian Male	-0.090**	-	-0.023	-	0.551	-
	(0.028)		(0.057)		(0.911)	
Indian Female	-0.059	-	0.013	-	1.601	-
	(0.036)		(0.061)		(1.580)	
Chinese Male	-0.055*	-	-0.036	-	-0.235	-
	(0.027)		(0.059)		(0.671)	
Chinese Female	-0.037	-	0.017	-	0.909	-
	(0.035)		(0.053)		(0.785)	
p-value (F test)	0.477	-	0.064	-	0.308	-
R^2	0.049	0.044	0.105	0.093	0.212	0.207

Notes: + $p < 0.10$, * $p < 0.05$ ** $p < 0.01$. All analyses condition on course, comment, and sequence fixed effects. The p-value refers to an F test for the joint equivalence of the effects associated with the 7 non-white male poster identities. Standard errors, presented in parentheses, are clustered at the course level. The sample size is 992 (i.e., 8 comments posted in each of 124 courses).

Table 3 - The Estimated Effects of a White Male Student Identity on Instructor and Peer Responses by Instructor, Course, and Comment Traits

	Dependent Variable			
Sample Construction	Instructor Replied	Student Replied	Number of Student Replies	Sample Size
Full Sample	0.058*	-0.021	-0.590	992
	(0.025)	(0.041)	(0.495)	
White Male Instructor	0.075*	-0.009	-0.611	576
	(0.037)	(0.055)	(0.566)	
Non-White Male Instructor	0.049	-0.026	-0.253	416
	(0.035)	(0.065)	(0.776)	
STEM	0.048	-0.031	-1.101	560
	(0.037)	(0.062)	(0.726)	
Non-STEM	0.043	0.008	0.361	432
	(0.031)	(0.065)	(0.661)	
Completion-Focused Comment	0.025	-0.015	0.035	433
	(0.029)	(0.068)	(0.341)	
Advising/Social Comment	0.060+	0.01	-0.101	559
	(0.033)	(0.054)	(0.925)	

Notes: + $p < 0.10$, * $p < 0.05$ ** $p < 0.01$. Each cell reports the estimated effect of a white-male poster identity relative to all other poster identities conditional on course, comment, and sequence fixed effects. Standard errors, presented in parentheses, are clustered at the course level. White-male instructor courses include single instructor courses taught by a white male and multiple instructor courses taught exclusively by white males. See Appendix Table A1 for comment catgorizations.

Table 4 - The Estimated Effects of Student Identities on Race and Gender-Congruent Peer Responses

	Dependent Variables	
Independent Variable	Student Replied	Number of Student Replies
White	0.059+	0.286
	(0.035)	(0.363)
Black	-0.007	0.000
	(0.014)	(0.022)
Indian	0.042+	0.091+
	(0.022)	(0.055)
Chinese	-0.009	-0.005
	(0.015)	(0.021)
Female	0.045+	0.375+
	(0.026)	(0.201)
White Male	-0.031	-0.095
	(0.041)	(0.192)
White Female	0.103**	0.504
	(0.038)	(0.321)
Black Male	-0.011	-0.022
	(0.015)	(0.018)
Black Female	0.027	0.043
	(0.018)	(0.027)
Indian Male	0.007	0.043
	(0.026)	(0.077)
Indian Female	-0.005	0.009
	(0.013)	(0.027)
Chinese Male	-0.003	-0.001
	(0.015)	(0.020)
Chinese Female	0.005	0.006
	(0.014)	(0.017)

Notes: + $p < 0.10$, * $p < 0.05$ ** $p < 0.01$. Each cell reports the estimated effect of the poster identity from a unique regression in which the dependent variable is a reply (or the number of replies) from peers with the poster's race and/or gender identity. We identified the race and gender of student peers for 64 percent of repliers (see text for details). All analyses condition on course, comment, and sequence fixed effects. Standard errors, presented in parentheses, are clustered at the course level. The sample size is 992 (i.e., 8 comments posted in each of 124 courses).

Appendix A - Categorization of Experimental Comments

We selected and posted a variety of comments intended to elicit different responses from discussion forum participants. We developed these comments by undertaking a pilot study in which we observed and curated the actual comments made in courses. We provide the comments used in our experiment below, although the wording has been slightly adjusted to protect the anonymity of course participants. The table below reflects the division of comments into two categories: completion-focused comments which focused on information necessary for the poster to successfully complete the course and were designed to elicit a response and advising/social comments which were less focused on successfully completing the course. We acknowledge that our classification is not the only way to divide our comments, so we examined several alterative categorizations of these course comments. The numbers in the second column represent four alternative categorizations: (1) a small reorganization of our preferred classification, (2) questions for the instructor versus questions for other students, (3) questions whose answer would be helpful to other students versus questions whose answer is only helpful for the asker, and (4) questions that are related to the content of the course versus those that are not related to content. For each of the alternate groupings, the number represents that that comment would switch to the other group. All groupings produced results that were qualitatively similar to those reported in Table 3.

Appendix Table A1 –Comment Classifications

Completion-Focused Comments	*Reclassifications*
I joined this class late and am wondering if I missed anything that is important.	3
Are there links to other resources that could be helpful for the lessons?	3
I am putting off watching the lectures. Does anyone have any tips to help me not procrastinate?	2, 3, 4
Should I watch the videos all at once or one by one? What do others do?	2, 3, 4
I'm finding the lectures difficult to follow. Anyone else?	3, 4
I haven't watched all of the lectures. I don't think I will be able to catch up -- what is the best lecture for me to watch?	2, 3, 4
How should I complete the assignments? Does anyone have tips on how to do them well?	3, 4
What's the minimum percentage I have to get to pass?	4
How do I submit assignments? Can someone please explain this to me?	4
Do we just have to watch the lecture videos? Is there anything else we have to do?	4
Are the lectures the only homework assignments? Is there anything else?	4
How do I find out how well I am doing in this class?	4
What kinds of things do I need to know to do well in this class?	

Advising/Social Comments	
Does anyone use this course material for their job?	4
Is this class harder or easier than other classes in this field?	4
Anyone have any ideas on courses that would be good to take after this one?	4
What is the goal of this class? Is it mostly theoretical or is it also practical?	2, 3, 4
Do people like this class? I am not sure I can finish it, but I might take it later. Is it worth it?	
I am learning lots from this class, even though it is a lot of work. Does anyone else feel this way?	
I am falling behind in this course. How is the workload?	1
Where are people in this class from?	
Are you taking this class for fun? Are you a student or are you working?	

I am struggling in this class. Does anyone else find it to be hard?	1
I am feeling more confident about this class, even though I struggled at first. Does anyone else feel this way?	
This class is challenging, and I am really enjoying the challenge!	
This class isn't as hard as I expected! I am enjoying it.	
I don't find all of the lectures to be that helpful.	
I am just starting week two of the class. Where are other people?	
This class is great. It is perfect timing for me!	
I don't have any prior experience. Will I do okay? What are the backgrounds of other people in the class?	1
Do you think I should put this class on my resume?	2
What do I need to do to unenroll from the class?	2

Appendix Table B1 - Testing the Balance of Student Identities by Comment and Comment Order

Student Identity	p-value
White Male	0.1131
White Female	0.8776
Indian Male	0.6817
Indian Female	0.9985
Black Male	0.6890
Black Female	0.9042
Chinese Male	0.5891
Chinese Female	0.0065

Notes: Each row is based on a separate regression in which the race-gender profile is regressed on indicators for comment and comment order. The p-value is based on an F-test of the joint significance of comment and comment-order fixed effects. Each regression also conditions on course fixed effects.

Appendix C – Multiple Comparisons

One possible concern with our main confirmatory findings in Table 2 is that the results may be an artifact of conducting "multiple comparisons." For example, our main family of estimates in Table 2 contains the results of 21 different statistical tests (i.e., assessing 7 point estimates for each of 3 outcomes). This approach resulted in 4 point estimates that are statistically significant at the conventional 95 percent level (and an additional statistically significant finding at the 90 percent level). If 5 percent of these inferences were in fact Type I errors, we would expect only 1 of these statistically significant findings to be a false positive (i.e., 0.05 * 21). To engage this concern more formally, we implemented the widely used procedure developed by Benjamini and Hochberg (BH; 1995). A key parameter in the BH procedure is the choice of a "false discovery rate" (FDR), the share of statistically significant findings (i.e., "discoveries") one is willing to accept as false positives.[1] Assuming an FDR of 0.10, we find that one of the four discoveries (i.e., the negative effects of an Indian male identity on the probability of an instructor response) remains statistically significant. Similarly, when we apply the BH procedure to the family of estimates that compare white males to all other identities (i.e., our other three columns in Table 2), our finding that white males were substantially more likely to receive an instructor response remains statistically significant. Overall, these results suggest that our main findings regarding instructor bias cannot be dismissed as an artifact of multiple comparisons.

[1] Efron (2012) notes that 0.10 is a popular choice. McDonald (2014) notes that "Sometimes people use a false discovery rate of 0.05, probably because of confusion about the difference between false discovery rate and probability of a false positive when the null is true; a false discovery rate of 0.05 is probably too low for many experiments."