

**Experimental Evidence on Teachers' Racial Bias in Student Evaluation: The Role of
Grading Scales**

David M. Quinn
Rossier School of Education
University of Southern California
quinnd@usc.edu

Quinn, D.M. (2020). Experimental Evidence on Teachers' Racial Bias in Student Evaluation: The Role of Grading Scales. *Educational Evaluation and Policy Analysis*.
doi:10.3102/0162373720932188

Acknowledgments: Funding was provided by a James H. Zumberge Individual Research Award from the University of Southern California. I am grateful to Drishti Saxena for indispensable research assistance, Daphna Oyserman for study design feedback, and to Amadou Diallo for assistance creating the writing sample used in the experimental manipulation.

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Abstract

A vast research literature documents racial bias in teachers' evaluations of students. Theory suggests bias may be larger on grading scales with vague or overly-general criteria versus scales with clearly-specified criteria, raising the possibility that well-designed grading policies may mitigate bias. This study offers relevant evidence through a randomized web-based experiment with 1,549 teachers. On a vague grade-level evaluation scale, teachers rated a student writing sample lower when it was randomly signaled to have a Black author, versus a White author. However, there was no evidence of racial bias when teachers used a rubric with more clearly-defined evaluation criteria. Contrary to expectation, I found no evidence that the magnitude of grading bias depends on teachers' implicit or explicit racial attitudes.

Key words: teacher bias; racial bias; implicit bias; grading rubric; randomized experiment

Experimental Evidence on Teachers' Racial Bias in Student Evaluation: The Role of Grading Scales

A vast research literature shows that teachers give racially biased evaluations of student work (for reviews, see Ferguson, 2003; Malouff & Thorsteinsson, 2016; Tenenbaum & Ruck, 2007). Downwardly biased evaluations can lead to actual reductions in student learning through self-fulfilling prophecies, or teacher expectancy effects (e.g., Ferguson, 2003). Such effects may be far-reaching, given that students' future teachers base their expectations in part on the biased evaluations of previous teachers. Biased evaluations may also produce stereotype threat (Steele, 2011), which negatively affects students' short-term performance and their learning over the longer term (Taylor & Walton, 2011). Furthermore, when students detect bias from their teachers, they are unlikely to develop trusting relationships with those teachers and may disengage from that class, or – over time – school more generally (Rangvid, 2018; Woodcock, Hernandez, Estrada, & Schultz, 2012).

Theory suggests that the magnitude of evaluation bias may depend on characteristics of the evaluation tool (Malouff & Thorsteinsson, 2016; Payne & Vuletich, 2018; Uhlmann & Cohen, 2005). When evaluation criteria are subjective or ambiguous, teachers' implicit or explicit stereotypes have greater potential to influence their grading. When evaluation criteria are clear and specific, teachers' judgments may be less susceptible to bias (Uhlmann & Cohen, 2005). The adoption and promotion of clear evaluation criteria may therefore be a simple policy lever for instructional leaders aiming to reduce racial bias in student evaluations. However, experimental research comparing bias across different scoring metrics is lacking.

In the present study, I replicate and extend experimental work on teacher bias in grading. I replicate - in a US setting – prior research from outside the US showing racial/ethnic bias in

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

teachers' evaluations of student work when performance criteria are relatively vague. I extend this work by studying whether such evaluation bias is present on a clear and specific criterion-referenced rating scale. I also study whether the magnitude of evaluation bias differs by teachers' implicit and explicit racial attitudes. In exploratory analyses, I examine whether bias differs by teacher demographics (i.e., racial match effects) or the racial demographics of teachers' school settings.

Background

Documenting Evaluation Bias

Researchers have documented racial bias in teachers' evaluations of students through observational, experimental, and quasi-experimental designs. Tennenbaum and Ruck (2007) conducted a comprehensive meta-analysis of the early studies on teachers' racial/ethnic biases for social and academic evaluations. The experimental studies they reviewed typically elicited teachers' ratings of students through vignettes or student work samples accompanied by photographs. In observational studies, teachers would often rate their actual students, then researchers would estimate bias by comparing teachers' covariate-adjusted ratings across social groups. The meta-analysis showed an average bias in favor of White students over Black students of $d=.25$ (30 studies), White students over Latinx students of $d=.46$ (6 studies) and Asian students over White students of $d=-.17$ (3 studies). Across the experimental studies, most manipulation methods (photo, audio/visual, simulated teaching) showed significant bias in favor of White students ($d= .21$ to $.51$), though vignette studies were an exception (with teachers showing average bias against White students of $d=-.10$ [13 studies]). The authors speculated this could be due to the vignettes not being realistic enough to trigger differential evaluations.

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

A more recent meta-analysis of the experimental research on grading bias by Malouff and Thorsteinsson (2016) examined studies of bias based on student race/ethnicity, gender, physical attractiveness, and disability status. Across 23 studies from 20 research articles, the authors found an average grading bias of $g=.36$. Seven of these studies (from 5 reports) examined racial or ethnic bias, with an average effect of $g=.26$. Only one of these studies took place in the US, which was a small, under-powered dissertation testing for Black/White bias (Gerritson, 2013). One goal of the present work is therefore to replicate previous bias research with US teachers in regards to Black/White bias.

In the more recent research on grading bias, scholars from outside of the US have used quasi-experimental methods to test for evaluator bias. Although the generalizability of this work to racial bias in American contexts is uncertain, the findings raise important questions in need of investigation domestically. Several of these studies have estimated gender bias in grading by comparing gender differences in scores on exams that were scored anonymously (i.e., the grader was not aware of the student's identity) to gender differences on (often separate) exams that were scored with knowledge of the student's identity (Falch & Naper, 2013; Hinnerich, Hoglin, & Johannesson, 2011; Lavy, 2008; Protivinsky & Munich, 2018; Rangvid, 2017; Terrier, 2016). Results from these studies have been mixed, with some finding teachers favoring females in math and reading (Falch & Naper, 2013; Lavy, 2008; Protivinsky & Munich, 2018), some finding teachers favoring females in math but not reading (Terrier, 2016), some finding teachers favoring males in math but not reading (Lavy & Sand, 2015), and others showing no bias at all (Hinnerich et al., 2011). Direction of the bias aside, longitudinal studies using this approach have suggested that teacher bias has long-term effects on students' future exam performance,

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

course-taking choices, and field of study (Lavy & Megalokonomou, 2019; Lavy & Sand, 2015; Terrier, 2016).

In recent experimental studies, researchers have asked teachers to score student work samples that were randomly assigned student names signaling different gender and ethnic identities. Two such studies found that German teachers scored essays more favorably when purportedly written by an ethnic German student compared to when purportedly written by a student of Turkish descent (Bonefeld & Dickhauser, 2018; Sprietsma, 2013). In India, teachers discriminated against lower caste students (by .03 to .09 SD) and high-performing girls (Hanna & Linden, 2009). However, a Dutch study using this method did not find bias in teachers' evaluations of students from Turkish or Moroccan backgrounds compared to ethnic Dutch students (van Ewijk, 2011). Methodologically, these studies offer evidence with strong internal validity regarding the causal effect of students' identities on teachers' evaluations. Such designs can help us better understand the extent to which similar bias effects generalize to the US.

Implicit Stereotypes and Teachers' Evaluations

What policy tools might successfully mitigate grading bias? Devising effective policy solutions requires that we understand the mechanisms underlying biased grading. Teachers' biased evaluations could be driven by either their explicit or implicit racial attitudes. Some teachers hold explicit racial stereotypes – or stereotypes they consciously endorse - which are liable to impact their treatment of racially minoritized students (e.g., Farkas, 2003; Quinn, 2017). However, it is likely more common for teachers to hold implicit racial stereotypes that influence their evaluations (Warikoo, Sinclair, Fei, & Jacoby-Senghor, 2016; Chin, Quinn, Dhaliwal, & Lovison, forthcoming). An implicit stereotype is one that is not identifiable through introspection (Greenwald & Banaji, 1995). Implicit stereotypes can be automatically activated in

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

one's mind (Devine, 1989), leading to biased behaviors or judgments (Greenwald & Krieger, 2006). Thus, teachers can exhibit implicit bias even when they do not consciously endorse the implicit stereotype from which it stems (Devine, 1989).

Implicit racial stereotypes may lead teachers to rate work produced by a Black student less favorably compared with the rating they would have given the same work had it been produced by a White student. Work by a Black student can automatically call to teachers' minds the stereotype of African Americans as unintelligent. This stereotype can then, perhaps unbeknownst to the teacher, lead them to judge the work as being consistent with the stereotype. As such, we would expect teachers holding stronger implicit racial stereotypes to exhibit larger racial biases when evaluating students. I test this hypothesis in the present study.

Policy Options for Mitigating Bias

Two main categories of policy levers are available to education leaders aiming to mitigate racial bias in grading. One approach focuses on professional development that “de-programs” individuals' implicit attitudes through methods such as repeated exposure to counter-stereotypical examples. A meta-analysis (Forscher et al., 2018) of 494 studies showed such interventions can be effective at reducing measures of individuals' implicit attitudes, with an average effect size of $d=.30$. However, these reductions in negative implicit attitudes did not lead to behavioral changes (Forscher et al., 2018). Consequently, we might expect that attempts to combat grading bias by training teachers to unlearn their general implicit stereotypes would be ineffective.

Another approach focuses on policies that engineer circumstances to reduce the influence that implicit attitudes exert on peoples' behaviors or judgments. Given that implicit stereotypes reflect content that is readily accessible in one's mind at the moment, stereotypes are less likely

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

to influence decision-making when people have an opportunity to process information more carefully (Payne & Vuletic, 2018). Implicit stereotypes can be more influential when cognitive load is high (e.g., when people are distracted), when processing capacity is diminished (e.g., when people are fatigued or under stress), or when time is limited. Teachers and administrators may therefore be able to reduce bias in grades by ensuring that evaluations are conducted free of distractions and with sufficient time. Under these circumstances, teachers have available the necessary cognitive resources to assess work fairly. This makes them less likely to rely on stereotypes in place of the more taxing or time-consuming cognitive work of sober evaluation. Another obvious strategy (for which empirical evidence is available through the studies described above) is to employ anonymous grading. However, anonymous grading will not always be feasible, and the improvements in teachers' grading conditions described above would not eliminate all pathways through which stereotypes can influence evaluations.

Clarifying performance criteria. Other specific policies that may reduce the influence of teachers' biases on grading are those that articulate how educators should evaluate student work. Although current debates about grading policy do not typically invoke the issue of racial bias (Brookhart et al., 2016; O'Connor, Jung, & Reeves, 2018; Reeves, 2008), some major grading reforms – such as standards-based grading (SBG) and mastery grading - may nevertheless offer the benefit of mitigating bias.

At the core of SBG and mastery grading policies is the philosophy that performance criteria should be predetermined and clearly specified (Brookhart et al., 2016). With SBG, student work is compared to grade-level performance levels articulated through ordered categories (such as “below basic,” “basic,” “proficient,” “advanced”). Mastery grading similarly

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

establishes clear mastery standards, but evaluates student performance on a binary mastered/not mastered scale (Brookhart et al., 2016).

Theory on implicit bias suggests that predetermined and clearly specified standards may help reduce grading bias because vague evaluation criteria leave more room for teachers' implicit biases to influence their judgements. If teachers are evaluating student work and they are unsure what standard to compare the work to, implicit stereotypes can "fill in the blanks." Additionally, research suggests that people shift their evaluation criteria to match the qualifications of people from groups they prefer (Uhlmann & Cohen, 2005). When evaluation criteria are clearly defined beforehand, there is no opportunity for such criteria-shifting. As such, teachers may exhibit less bias when clear and specific evaluation standards are employed, versus when vague and general criteria are employed. For example, if teachers are asked to rate a piece of student writing on a scale of 1-10 where higher values simply indicate higher quality, teachers may shift their indicators of quality to match their biased expectations about which student groups would produce the higher quality writing. In contrast, if the dimensions of evaluation are predetermined through specific writing traits – and if criteria for specific performance levels are clearly articulated – teachers will be oriented toward the aspects of the student work that are relevant to their evaluation.

Despite the popularity of SBG and mastery grading among reformers, research suggests that US teachers typically have a great deal of autonomy in how they determine student grades (Brookhart et al., 2016). When formal grading policies do exist, they tend to involve broad strictures such prohibiting teachers from giving students zero credit on assignments (Walker, 2016), or policies that limit the extent to which nonachievement factors can impact students' grades (Cox, 2011). Furthermore, evidence suggests that the implementation of grading policies

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

varies widely across teachers (Cox, 2011), with many teachers lacking familiarity with the policies (Tierney, Simon, & Charland, 2011). Grading policy – implemented with a focus on teacher training and buy-in – deserves more attention as a potential tool for mitigating racially biased grading practices.

There is suggestive evidence that policies that clarify performance standards may help reduce grading bias. In the experimental and quasi-experimental studies discussed above, information is not always provided regarding whether explicit evaluation criteria were used by raters. In the cases in which rater discretion was explicitly noted or in which vague rating scales were described, grading bias against various social groups was found (Bonefeld & Dickhauser, 2018; Hanna & Linden, 2009; Sprietsma, 2013). In contrast, no evidence of bias was found in the one study that noted the use of explicit grading guidelines (Hinnerich et al., 2011). In their meta-analysis, Malouff and Thorsteinsson (2016) tested whether the use of grading rubrics moderated the magnitude of bias estimates across studies. Although effect sizes did not differ significantly depending on rubric use, results were suggestive: The average effect size across the 6 studies using rubrics was not significant and was smaller in magnitude than the average effect size across the 17 studies that did not use a rubric ($g=.24$ vs. $.39$). However, these estimates are noisy and cross-study confounds cannot be ruled out.

If some evaluation methods or metrics are less susceptible to bias than others, instructional leaders have a simple low-cost policy tool available for reducing bias in grading. Leaders at the school- and district-levels may be able to design evaluation procedures to minimize bias, provide professional development on these procedures, and even encourage their use through negotiating professional teaching standards. Teacher preparation programs also may have a role to play in preparing and encouraging teachers to use evaluation tools or procedures.

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

At the state-level, procedures for scoring statewide writing tests may be designed to minimize bias. The first step is to produce evidence regarding the extent to which different evaluation metrics yield biased scores.

Summary and Study Hypotheses

Plentiful research has documented bias in teachers' evaluations of student work. However, very little recent experimental research has been conducted in the US on Black/White bias in grading. Theory suggests that less bias may be evident when clear criterion-referenced evaluation standards are employed, versus vague and general criteria. If biases are more likely to appear on some rating scales than others, adopting tools less susceptible to bias may be a simple policy lever for reducing biased grading. However, we lack experimental research on the extent to which bias exists across different measures. Finally, it will be beneficial to examine whether measures of teachers' implicit or explicit racial attitudes moderate grading bias. Such knowledge will be useful to teachers and policymakers who are working to eliminate the effects of bias on student evaluation.

In this study, I begin by replicating past experimental work on grading bias, but in a US context and with regard to Black/White bias. Most importantly, I hypothesized that: 1) teachers would show racially-biased evaluations of student writing when employing a vague grade-level rating scale, but 2) teachers would show less, or no, bias in evaluations when given a more specific criterion-referenced rating scale, and 3) grading bias would be larger among teachers holding stronger implicit stereotypes of Black students as lacking competence (as measured by an implicit association test), and with more explicit preference for European versus African Americans (as measured by feeling thermometers). I find evidence in favor of the first two hypotheses but not the third. Additionally, I conduct exploratory analyses examining whether

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

the magnitude of bias on the vague relative scale differed by characteristics of teachers or their schools.

Methods

Participants and Procedures

I conducted this web-based survey experiment by contracting with Qualtrics to recruit a national (though not nationally-representative) sample of US schoolteachers (respondents were compensated directly by Qualtrics). The target sample size was 800 teachers, yielding .80 power to detect a bias main effect of .20 SD (in the binary outcome metrics employed below, this yields .80 power to detect a treatment/control group difference of approximately .10 or .07 for control group proportions of .50 and .10, respectively). All panel participants who were preK-12 teachers were eligible, and Qualtrics terminated data collection when 810 participants had completed the entire survey. Of the 1,799 unique respondents who clicked the survey link, 163 were terminated immediately because they were not in fact teachers. The remaining 1,636 participants completed the experimental phase of the survey (writing evaluation task, demographic questionnaire, and explicit bias measure).¹ During administration of the implicit association test (described below), however, a large share of respondents abandoned the survey (hence the sample size discrepancies between the main effects analyses and the IAT moderation analyses). Of those who completed the entire survey, 706 produced valid IAT scores (with some participants' excessive speed preventing score calculation). Finally, a total of 87 respondents were dropped from the analyses because they did not self-identify as a current full-time preK-12 teacher (e.g., retired teachers, substitute teachers, teachers' aides, parents home-schooling their children). This resulted in an analytic sample of 1,549 for the main effect analyses and the analyses testing for moderation by explicit racial attitudes; 675 teachers were included in the

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

analyses testing for moderation by implicit stereotypes (see Appendix A for comparisons of the IAT analytic sample versus those not in the IAT analytic sample, as well as comparisons across experimental condition among teachers in the IAT analytic sample). The actual sample sizes yield .80 power to detect a main effect of .14 SD, or approximately .07 and .05 for treatment-control proportion differences when control group proportions are .50 and .10, respectively.

At the start of the survey, teachers were informed that the researcher was interested in learning how educators evaluate student writing. Teachers then answered questions about their teaching background (current position, years in the field of education) before being randomly assigned to receive one of two versions of a student writing sample that used different names to signal a Black or White student author (described below). Participants were required to respond to each item before proceeding to the next.

In Table 1, I present descriptive statistics by experimental condition and balance checks for the full analytic sample(s). Differences across conditions in pretreatment characteristics were not large (though some were statistically significant or marginally significant; see Table 1). The “Black author” group had slightly more female teachers (64% vs. 59%) and slightly more K-2 and grade 3-5 teachers. The sample was majority White (~ 69%); the modal teacher taught in a predominantly White school (~54%) and had been in the field of education for 7-10 years (~24%). For comparison, I include statistics for teachers nationally, as available.

<Insert Table 1>

Experimental Materials

Teachers were shown a scanned copy of a student writing sample purportedly written by a student in the fall of second grade in response to a prompt to write about their weekend. In the response, the student author refers to his brother by name, as well as his brother's friend. In one

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

condition, teachers randomly received a response in which the writer's brother's name is "Dashawn," signaling a Black author. In the other condition, teachers randomly receive a response in which the writer's brother's name is "Connor," signaling a White author (names taken from list of most racially distinct names; Levitt & Dunbar, 2005). In full, the student essay read, "I wose with my brother [*Dashawn/Connor*] and his frind [*Arin/Scot*] but it wose a graet day to be a boy at home..." (see Appendix B for actual experimental materials). Teachers then rated the writing on the scales described below, answered demographic questions, and completed measures of implicit and explicit racial attitudes.

As a subject area, writing is well-suited for a study of grading bias for two main reasons. Substantively, the subject area is of interest given that tools for evaluating student writing vary in their focus and specificity. Methodologically, the personal narrative lends itself well to signaling the author's racial identity in a relatively subtle way (as described above). This can help reduce demand effects, given that an explicit statement of the student's racial identity may arouse suspicion among research participants.

Measures

Writing evaluations. Teachers were first asked to rate the writing sample on a relative grade-level scale with options "far below grade level," "below grade level," "slightly below grade level," "at grade level," "slightly above grade level," "above grade-level," and "far above grade level." This scale represents the vague, general scale because performance criteria are not explicitly defined. As discussed below, I convert these responses to a binary "at or above grade level" scale for interpretability in my main analyses.

Teachers then rated the writing on a rubric with more clearly defined performance criteria. This item read, "Overall, where would you place this student's writing on the following

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

rubric for a personal narrative?" The rubric was comprised of the ratings "provides a well-elaborated recount of an event," "recounts an event with some detail," "attempts to recount an event," and "fails to recount an event." In my main analyses, I convert this to a binary scale of "recounts an event with some detail" or better (see Appendix C for ordered logistic regression models that use the original numeric versions of the scales; all results are robust). The rubric appeared after the grade-level scale (and on a separate screen without the option to return to the earlier screen) to ensure that teachers' ratings on the grade-level scale were not influenced by the criteria in the rubric.

Substantively, these evaluation measures differ in two important respects. The grade-level scale is general in the sense that it does not specify what dimension(s) the rater should consider (e.g., grammar, spelling, creativity, organization, etc.). It also does not clearly specify the gradations among scale points (e.g., how should a teacher determine whether the writing is "slightly above grade-level" versus "above grade-level"?). These ambiguities are hypothesized to leave the grade-level scale more susceptible to bias. In contrast, the rubric specifies the evaluation domain of interest (how well the writer recounts an event) and provides more specific scale point descriptors to guide teachers in their rating choices. The two measures also differ in the number of possible scale points, a matter I return to in the Discussion section. When interpreting the variation in bias estimates across these two measures, the totality of these differences should be kept in mind.

By using single-item evaluation outcomes, I follow the convention in the experimental literature on grading bias (Bonefeld & Dickhauser, 2018; Hinnerich, Hoglin, & Johannesson, 2011; Rangvid, 2017; Sprietsma, 2013; van Ewijk, 2011) and improve ecological validity by mimicking the grading process as it often occurs in real-world settings. However, a drawback of

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

these measures is that I am unable to calculate reliability statistics for the sample. I return to this consideration in the Discussion section.

Racial attitudes. In this study, collecting data on respondents' racial attitudes presented a challenge. If these measures are administered before teachers see the writing sample, the act of completing the racial attitude measures may produce demand effects that influence teachers' ratings of the writing sample. If respondents complete the racial attitude measures after viewing the writing sample, the writing sample may impact their racial attitude scores. I opted for the second sequence to prevent contamination of my first two hypothesis tests, viewing this as less damaging to the experiment overall (furthermore, experimental effects on racial attitudes were tested for and showed null results).

Implicit competence stereotypes. To measure teachers' implicit stereotypes of Black students, I adapted the traditional implicit association test (IAT) using the iatgen online software (Carpenter et al., 2018). The IAT is a timed computerized classification test that assesses “the strength of association between a target category and two poles of an attribute dimension” (Nosek & Banaji, 2001, p. 627). The traditional race IAT has been validated and found to be predictive (albeit weakly) of various biased behaviors (Greenwald, Poehlman, Uhlmann, & Banaji, 2009; but see Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013 for a different take on the evidence). In my adapted IAT, the target category is race (African American/European American) and the two poles are competence and incompetence. A positive IAT “*d*-score” indicates more pro-White bias (i.e., stronger implicit stereotypes of White students as more competent than Black students), a negative score represents pro-Black bias, and zero represents neutrality. My adapted competence IAT demonstrated internal consistency (based on split-half

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

with Spearman-Brown correction) of .86 (see Appendix D and Quinn [forthcoming] for additional detail on this measure, its development, and validity evidence).

Explicit attitudes. I administered traditional feeling thermometers (Nelson, 2008) in which teachers rated how warm or cold they feel toward Black and White Americans. The items read, “How cold or warm do you feel toward African Americans [European Americans]?” A 1-10 scale was shown with 1 representing “very cold” and 10 representing “very warm.” I created a measure of explicit bias by calculating the difference, for each individual, in their rating of White and Black Americans (such that positive scores indicate a preference for White Americans, negative scores a preference for Black Americans). Again, while such single-item measures preclude internal consistency estimates, feeling thermometers such as these are widely used in psychology, sociology, and political science (Nelson, 2008; Xu, Nosek, & Greenwald, 2014).

Analytic plan

In their original form, the rating scales used in this study are ordinal. To improve interpretability, I dichotomize these scales and fit linear probability models (again, Appendix C includes results from ordered logistic regression models that use the original full scales as outcomes; all conclusions are robust). My models take the form:

$$P(y_i = 1) = \beta_0 + \beta_1 BLACKAUTHOR_i + \sum \beta_j X_j$$

In one set of models, y_i is a 0/1 indicator for whether respondent i rated the writing sample as being on grade-level or above (versus below grade-level). In separate models, y_i is a 0/1 indicator for whether respondent i rated the writing as “recounts an event with some detail,” or better. $BLACKAUTHOR_i$ is a binary indicator for whether the teacher was randomly assigned to the “Dashawn” version of the writing sample and $\sum \beta_j X_j$ is a set of fixed teacher characteristics

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

included to improve precision of the bias estimates while controlling for chance imbalances across conditions (unadjusted estimates are also shown). These controls include teacher race/ethnicity, gender, grade-level assignment, experience level, and whether the student body at their school is primarily Black. In these models, β_1 is the coefficient of interest, representing the extent to which teachers' evaluations are racially biased, on average. I hypothesized this coefficient would be negatively-signed and statistically significant for the vague grade-level outcome measure and smaller in magnitude or statistically zero for the criterion-based personal narrative rubric. All models are fit with heteroskedasticity-robust standard errors.

To explore variation in racial bias, I break the full sample down into several subgroups and fit models separately by teacher race/ethnicity, gender, and the racial demographics of their school. These analyses should be viewed as exploratory, conducted for the purpose of generating (rather than confirming) hypotheses.

To test whether the magnitude of racial bias differed by teachers' implicit or explicit racial attitudes, I fit a series of models that include the main effect of one of the attitude measures along with its interaction with the $BLACKAUTHOR_i$ indicator. I hypothesized that in the grade-level rating models that use either the competence IAT or the explicit feeling thermometer as a moderator, the main effect of $BLACKAUTHOR_i$ would be close to zero (given that a value of 0 on these measures indicates no bias). However, I expected that the interaction term would be negative and significant, indicating that teachers with stronger implicit or explicit anti-Black (or pro-White) attitudes exhibit stronger bias against Black students on the grade-level rating scale.

See Appendix E for estimates from logistic regression models (all results are robust).

Results

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Racial Bias on Grade-level and Rubric Scales

In the top panel of Table 2, I present main effect estimates (i.e., $\hat{\beta}_1$) of student's implied racial identity on teachers' grade-level and rubric ratings for the full sample. For each outcome, I include unadjusted estimates along with estimates that adjust for the set of controls described above. I also display the adjusted outcome means for the "Connor" group. In the second column of estimates, we see evidence of racial bias in teachers' evaluations on the vague grade-level scale (after controlling for teacher characteristics).

<Insert Table 2>

Teachers who were shown the *Dashawn* version were 4.7 percentage points less likely to rate the writing as being on grade-level or above compared with teachers shown the *Connor* version (with 35% of respondents rating the White version as grade-level or above [adjusted mean column]). Consistent with my hypothesis, teachers gave essentially identical ratings to the Black and White authors on the more explicit rubric (right side panel). Approximately 37% of teachers (adjusted) rated the "Connor" and "Dashawn" version of the prompt as recounting an event with "some detail" or better.

Theoretically, bias may be stronger among teachers less familiar with appropriate expectations for students of this age. Teachers of lower-elementary grades will presumably have more useful background knowledge to draw from when evaluating the writing sample, while other teachers may be more likely to allow stereotypes to "fill in the blanks" where their expertise is lacking. One limitation of this sample, therefore, is that it includes teachers from across all grade-levels (this choice was made for practical reasons, given the cost of setting grade-level qualifiers for the sample). I therefore estimated bias separately for K-2 teachers and all other teachers (first two rows of "Subgroup Effects" panel of Table 2). The bias estimate

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

among K-2 teachers (approximately 10 percentage points) was larger in magnitude than the estimate for non-K-2 teachers (though not significant due to the small subgroup sample size, $n=227$; $p=.13$). In a linear probability model interacting each grade-level band indicator with the “Black Author” indicator (along with grade-level main effects), I failed to reject the null of equal bias across all grade-level groups ($p=.37$). These results provide reassuring evidence that the full sample bias estimates were not being driven by teachers of more advanced grade levels.

Research shows that teacher biases can arise as demographic match effects, such that teachers show preference for students with identities similar to their own (e.g., Gershenson, Holt, & Papageorge, 2016). I therefore include in Table 2 exploratory subgroup analyses examining whether the bias on the relative grade-level scale differs by teacher gender or race/ethnicity.

In the “Males” and “Females” rows in the “Bias Estimates by Subgroup” panel of Table 2, we see that the main effects for the full sample were driven by female teachers. Females were 7 percentage points (adjusted) less likely to rate the Black author as being on grade-level, but the effect for males was small and non-significant ($b=.002$, *n.s.*; in a test of the null hypothesis of equal bias for males and females, $p=.07$). Female teachers may be more likely than male teachers to show bias against Black males.

White teachers exhibited the largest bias against the Black student author. White teachers were approximately 8 percentage points less likely to rate the Black author’s writing as being grade-level or above, compared to the White author’s. This bias among White teachers was significantly different ($p=.03$, not shown) from the bias among all other teachers, who collectively showed a non-significant preference for the Black student author ($b=.03$, $p=.50$, not shown). (White teachers were also the only racial/ethnic group with a statistically significant bias, though note that other subgroups had substantially smaller sample sizes). Moreover, White

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

teachers were more likely than others to rate the White student's response as being on grade-level or above, suggesting that part of their bias may be due to an in-group preference. The significance level of these bias estimates for female teachers and White teachers are robust across all specifications. (See Appendix F for additional exploratory analyses broken down for gender-by-race teacher subgroups; again, these subgroup estimates should be interpreted cautiously given the small sample sizes).

In the right-side panel of Table 2, we see that no bias estimate was statistically significant on the personal narrative rubric, and most estimates were small in magnitude.

Given that some teachers have more Black students than others, it is worth examining the extent to which teachers' biases vary across schools with different racial demographics. In Table 3, I report the bias estimates from the relative grade-level measure (left panel) and the rubric (right panel), broken down by the racial make-up of the teachers' schools. Here, we see that bias was largest for teachers in racially diverse schools, at 13 percentage points ($p < .05$). Effects were small and not significant for teachers in primarily Black ($b = .01, n.s.$) and primarily Latinx ($b = -.01, n.s.$) schools. Among teachers in primarily White schools, bias was not statistically significant, though the magnitude matched that of the overall sample ($b = -.047, n.s.$). In no school type was there evidence of bias on the personal narrative rubric.

<Insert Table 3>

Interactions with Racial Attitudes

As discussed above, the magnitude of bias in teachers' evaluations may depend on teachers' racial attitudes. Teachers with stronger implicit stereotypes of Black students as less competent than White students, or with less explicit warmth toward African Americans versus European Americans, may show more bias on the grade-level evaluation measure. Recall that

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

the racial attitude measures were administered after teachers rated the writing sample. As seen in Table 1, there is no evidence that the writing sample version affected any of the attitude measures. I therefore proceed with using attitudes as moderator variables.

In Table 4, I present the results from linear probability models that test whether experimental condition interacts with measures of implicit or explicit racial attitudes (again, sample sizes for analyses with the IAT are substantially reduced due to participant drop-off in these phases of the survey). Columns 1-2 show models with the vague grade-level rating outcome, and columns 3-4 show models with the more specific rubric rating outcome. In each column, the treatment indicator interacts with a different implicit or explicit bias measure. As can be seen across models, in no case does the magnitude of the bias differ significantly by teachers' implicit or explicit racial attitudes.

<Insert Table 4>

Discussion

In this study, I found evidence of racial bias in teachers' evaluations of student writing when scored using a vague relative grade-level rating scale. However, there was no evidence of bias when teachers scored the writing using a more descriptive rubric with absolute criteria. These findings are consistent with theory from scholars of implicit bias (Payne & Vuletich, 2018; Uhlmann & Cohen, 2005). Teachers' stereotypes may have more influence on their evaluations when they are not given clear, specific criteria on which to rate student work. In contrast, teachers may be less likely to draw on their stereotypes when they have less discretion over the criteria for evaluating students.

I did not find evidence that teachers' implicit or explicit racial attitudes moderated their biased evaluations. Given the sample size for the implicit bias moderation analysis (n=675) and

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

the internal consistency of this IAT (.86), power is .80 to detect a bias*treatment interaction of approximately -.11 SD (when overall “grade-level or above” proportion is .40). One possibility, then, is that this study was under-powered to detect the true moderation effect. Another possible explanation is that the strength of a teacher’s implicit racial stereotypes does not affect the likelihood that they will evaluate student work in a racially biased manner. Knowing that explicit racial attitudes often diverge from implicitly-measured attitudes, skeptics have argued that perhaps explicit attitudes dominate in determining behavior, making implicitly-measured attitudes less of a behavioral concern (Oswald et al., 2013). Some divergence in explicit and implicit attitudes was observed in the present study; while the IAT showed, on average, a significant implicit association of White students as being more competent than Black students ($d=.41$), the explicit measure showed a small but non-significant preference for White Americans (.068 points on the feeling thermometer, or .047 SD, $p=.068$ for $H_0: \mu_{expbias} = 0$). However, if in fact teachers were summoning explicit attitudes to override an initial implicit instinct to rate the “Dashawn” writing prompt lower on the grade-level scale, the experimentally observed grading bias suggests they were not entirely successful. The influence of implicit attitudes on grading may therefore have been dampened but not entirely eliminated (which would be consistent with the negatively-signed but small and non-significant interaction term). Perhaps the magnitude (and significance) of the interaction would have been more pronounced had the experiment imposed time constraints or increased cognitive load.

This study provides direct evidence regarding grading bias as it manifests for a particular writing sample on two particular rating scales. Conceptual replications will be useful in producing evidence as to whether bias varies across other rating scales or other student work samples. Do the present results suggest that, more generally, using explicit grading criteria will

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

help mitigate grading bias? Or is there something unique about this rubric, this writing sample, or the combination of the two? We might expect that rubrics will be more effective at mitigating bias as the clarity of their performance criteria increases. Similarly, the more clearly a particular work sample meets a given set of criteria, the greater the bias-mitigating effect might be.

We should also consider whether bias may differ depending on the academic subject or the nature of the work being evaluated. The evaluation of student writing is likely more subjective than other types of evaluation, such as whether a student arrived at the correct answer to a math problem. Indeed, if my proposed explanation for the observed differences across the evaluation metrics in the present study is correct, we might expect less bias on teachers' grading of math problems as correct or incorrect.

Exploratory analyses suggest that Black/White grading bias toward male students on the relative grade-level measure may be stronger among White teachers and female teachers. In fact, White, Latina, and Black female teachers all showed similar estimates of bias (though the bias was only statistically significant for the White subgroup, which had the largest sample size; see Appendix F). This raises the question of how student/teacher match on race and gender might operate together when it comes to biased evaluations. We know from past research that teachers sometimes rate same-race students more favorably, and the results of the present study suggest that some demographic match effects may operate differently depending on other demographic traits. Are teachers less likely to exhibit racial bias against a student if the student shares their gender? Again, these results should be taken as hypothesis-generating rather than as confirmatory.

It is somewhat reassuring to see that teachers in primarily Black schools showed no evidence of racial bias in their evaluations, given that these teachers interact with many Black

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

students. Teachers at racially diverse schools showed the largest bias, suggesting that Black students at diverse schools may be especially at risk of receiving biased evaluations. Given that these analyses are exploratory, and that this sample is not nationally-representative (though national in scope), we cannot know whether this finding reflects patterns in the broader population. Taking the finding at face value, however, it is consistent with some theories on implicit bias. On average, Black students score lower on standardized tests than White students, and this holds within schools (e.g., Fryer & Levitt, 2004; Quinn, 2015; Quinn & Cooc, 2015). On average, then, teachers in racially diverse schools will more regularly encounter inter-group comparisons in which White students perform higher than Black students. This may lead these teachers to develop stronger implicit (or explicit) stereotypes of Black students as less competent than White students. When this stereotype is more accessible in one's mind, it can be more influential on one's judgments.

Policy Implications

As discussed above, scholars have focused on two distinct approaches for mitigating the effects of negative implicit attitudes: training programs that aim to reduce people's general implicit associations, versus efforts that engineer circumstances to reduce the impact that people's implicit stereotypes can have on their behaviors or judgments. The present study lends support to one form of the latter strategy in the context of teacher education and development. Given that teachers showed no bias when using explicit evaluation criteria, education leaders and teacher education programs may be more effective at reducing grading bias if they focus on implementing policies that establish predetermined and clearly-defined grading criteria.

I also found evidence that the overall bias on the vague grade-level scale was driven by White teachers, whose bias estimate was significantly different from the non-significant bias

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

estimate among all other teachers. This bias may have been driven by White teachers showing in-group preference toward White students. This finding aligns with calls to diversify the teaching force. Recent research has shown that although the share of teachers of color has grown in recent years, this growth has not kept pace with the increase in the share of students of color (Hansen & Quintero, 2019). The present findings suggest that the relative overrepresentation of White teachers may disadvantage Black students (potentially through White teachers showing undue preference for White students). In addition to promoting strategies effective in reducing teachers' evaluation biases, policies aimed at recruiting and retaining teachers of color may also help reduce the frequency of Black students' experiences with biased evaluations.

In the present study, teachers were simply presented with a grading rubric without any training or norming examples. Previous evidence has suggested that the benefits offered by explicit rubrics for increasing score reliability may require that teachers receive training on the rubric (Rezaei & Lovorn, 2010). When it comes to the potential for clear evaluation rubrics to reduce bias, such training may not be necessary – at least for simple writing responses in the early grades, evaluated with straightforward rubrics. For cases in which teachers are evaluating more extensive writing artifacts and employing more complex rubrics, prior training may be necessary before teachers can understand the criteria deeply enough to apply them without bias. Additionally, policies aiming to specify grading standards or practices face the same challenge faced by many education policies: the challenge of penetrating the classroom to actually change teacher practice (Weick, 1976). Such policies will be more likely to influence practice if they are accompanied by effective teacher training or coaching (e.g., Kisa & Correnti, 2015).

Limitations and Future Work

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

As noted earlier, the use of single-item evaluation measures in this study follows the convention of this research area and strengthens ecological validity. At the same time, it prevents the calculation of reliability statistics for the outcomes. Given that outcome measures with lower reliability yield lower statistical power, one potential concern would be that the differences in experimental effects across outcomes found here could be due to differential reliability of the measures. However, this would require that the vague grade-level measure have higher reliability than the more clearly-defined rubric, which contradicts past research (e.g., Jonsson & Svingby, 2007) and intuition. As additional reassurance, the difference in significance levels across outcomes is driven primarily by the differences in the magnitude of the estimates rather than the differences in precision, with the group difference being close to zero for the rubric outcome ($b = -.008$ for rubric, versus $b = -.047$ for grade-level rating).

Another possibility is that the ordering of the writing evaluations affects the scores teachers give. In this study, teachers rated the writing sample on the grade-level scale before rating it on the rubric. The purpose of this ordering was to ensure the criteria in the well-defined rubric did not influence the criteria that teachers had in mind when applying the vague grade-level scale. However, it is possible that bias on the vague rating scale could be reduced by having teachers first rate the writing sample using the rubric. In other words, focusing teachers' attention to a clear set of criteria may have carry-over effects to reduce bias in evaluations more generally. We also cannot know the extent to which the number of scale points on each measure may have affected the appearance of bias. The four-point personal narrative rubric was taken directly from an actual writing rubric, while the choice of a 7-point grade-level scale was made to align with recommendations for developing bi-polar response scales (Gehlbach & Brinkworth,

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

2011). I cannot rule out the possibility that the number of scale points affects the amount of bias detected by the scale.

The generalizability of these findings to classroom settings is unknown. The extent to which bias appears in teachers' evaluations of their own students – with whom they have relationships and of whom they have prior knowledge – may differ compared to this experimental setting. The present findings may be more generalizable to settings where raters are conducting anonymous review of essays in which the authors' identities may be signaled through context clues. Such settings would include the grading of state writing exams, or potentially SAT or GRE scoring.

If teachers are less likely to give biased evaluations of their own students (compared to students they do not know), then explicit rubrics may offer less benefit than the present study would suggest (at least with regards to the goal of mitigating bias). However, past research comparing anonymized and non-anonymized scoring methods has found evidence of teachers' gender bias directed toward their own students (e.g., Falch & Naper, 2013; Lavy & Sand, 2015; Terrier, 2016). There may therefore be reason to recommend absolute rubrics as a means to mitigate bias. Yet the present study does not offer direct evidence on whether rubrics would produce bias-reducing effects in such a setting. It is possible that teachers hold strong student-specific biases that absolute rubrics are less effective at overcoming. One potential way to study this in a field setting would be to randomly assign teachers to grade their own students' writing using either a vague scale or more explicit rubric, and then compare Black/White differences in scores on the two measures.

Conclusion

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Teachers' biased evaluations of student work may lead to a vicious cycle in which initial racially biased evaluations from a teacher cause lower future performance from students, which reinforces stereotypes held by teachers, which in turn leads to future bias in evaluations. This is a cycle over which teachers, school leaders, and district policies may be able to exert some influence through engineering evaluation procedures in bias-minimizing ways. The present study suggests that bias does not appear equally across all evaluation scales. The findings are consistent with the hypothesis that teachers exhibit less bias when they are given clear and specific grading criteria. By developing a deeper understanding of why some evaluation methods may be less likely to yield bias than others, we may be able to equip educators with a simple tool for mitigating bias: the thoughtful selection of evaluation measures.

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Notes

¹This phase of the survey also included items unrelated to the present line of inquiry (administered after this study's grading items), including implicit and explicit measures of the extent to which teachers view various student subgroups as warm or competent (Fiske, Cuddy, Glick, & Xu, 2002) and the measures reported in Quinn, Desruisseaux, & Nkansah-Amankra (2019).

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

References

- Bonefeld, M., & Dickhauser, O. (2018). (Biased) grading of students' performance: Students' names, performance level, and implicit attitudes. *Frontiers in Psychology, 9*, 1-13.
- Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F...& Welsh, M.E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*, 803-848.
- Carpenter, T., Pogacar, R., Pullig, C., Kouril, M., LaBouff, J., Aguilar, S. J., ... Chakroff, A. (2018, April 3). Conducting IAT Research within Online Surveys: A Procedure, Validation, and Open Source Tool. <https://doi.org/10.31234/osf.io/6xdyj>
- Chin, M.J., Quinn, D.M., Dhaliwal, T.K., & Lovison, V.S. (forthcoming). Bias in the Air: A Nationwide Exploration of Teachers' Implicit Racial Attitudes, Aggregate Bias, and Student Outcomes. *Educational Researcher*.
- Cox, K. B. (2011). Putting classroom grading on the table, a reform in progress. *American Secondary Education, 40*(1), 67-87.
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5-18.
- Falch, T., & Naper, L.N. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review, 36*, 12-25.
- 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*(6), 1119-1146.
- Ferguson, R. F. (2003). Teachers' perceptions and expectations and the black-white test score gap. *Urban Education, 38*, 460-507.

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

- Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6), 878.
- Forscher, P. S., Lai, C. K., Axt, J., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2018, July 13). A Meta-Analysis of Procedures to Change Implicit Measures. <https://doi.org/10.31234/osf.io/dv8tu>
- Fryer, R.G., & Levitt, S.D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics*, 86, 447-464.
- Gehlbach, H., & Brinkworth, M.E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, 15, 380-387.
- Gerritson, M. (2013). Rubrics as a mitigating instrument for bias in the grading of student writing (Doctoral dissertation, Walden University).
- 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.
- Greenwald, A.G., & Banaji, M.R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4-27.
- Greenwald, A.G., & Krieger, L.H. (2006). Implicit bias: Scientific foundations. *California Law Review*, 94, 945-967.
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17-41.
- Hanna, R., & Linden, L. (2009). Measuring discrimination in education. NBER Working Paper

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

No. 15057.

- Hansen, M., & Quintero, D. (2019). The diversity gap for public school teachers is actually growing across generations. *Brown Center Chalkboard, Brookings Institution*. Retrieved from: <https://www.brookings.edu/blog/brown-center-chalkboard/2019/03/07/the-diversity-gap-for-public-school-teachers-is-actually-growing-across-generations/>
- Hinnerich, B.T., Hoglin, E., & Johannesson, M. (2011). Are boys discriminated in Swedish high schools? *Economics of Education Review, 30*, 682-690.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130–144.
- Kisa, Z., & Correnti, R. (2015). Examining implementation fidelity in America's choice schools: A longitudinal analysis of changes in professional development associated with changes in teacher practice. *Educational Evaluation and Policy Analysis, 37*, 437-457.
- Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92*, 2083-2105.
- Lavy, V., & Megalokonomou, R. (2019). Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study. NBER Working Paper No. 26021.
- Lavy, V., & Sand, E. (2015). On the origins of gender human capital gaps: Short- and long-term consequences of teachers' stereotypical gender biases. NBER Working Paper No. 20909.
- Levitt, S. D., & Dubner, S. J. (2005). *Freakonomics: A rogue economist explores the hidden side of everything*. New York: William Morrow.
- Malouff, J.M., & Thorsteinsson, E.B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education, 60*, 245-256.

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

- Nelson, S.C. (2008). Feeling thermometer. In P.J. Lavrakas (Ed.), *Encyclopedia of survey research methods*, (pp. 275-277). Thousand Oaks: Sage Publications, Inc.
- Nosek, B.A. & Banaji, M.R. (2001) The go/no-go association task. *Social Cognition, 19*, 625-664.
- O'Connor, K., Jung, L.A., & Reeves, D.B. (2018). Gearing up for FAST grading and reporting. *Phi Delta Kappan, 99*(8), 67-71.
- 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.
- Payne, K.B., & Vuletich, H.A. (2018). Policy insights from advances in implicit bias research. *Policy Insights from the Behavior and Brain Sciences, 5*, 49-56.
- Protivinsky, T., & Munich, D. (2018). Gender bias in teachers' grading: What is in the grade. *Studies in Educational Evaluation, 59*, 141-149.
- Quinn, D. M. (2015). Kindergarten Black–White test score gaps: Re-examining the roles of socioeconomic status and school quality with new data. *Sociology of Education, 88*(2), 120-139.
- Quinn, D.M. (2017). Racial attitudes of preK-12 and postsecondary educators: Descriptive evidence from nationally representative data. *Educational Researcher, 46*, 397-411.
- Quinn, D. M., & Cooc, N. (2015). Science achievement gaps by gender and race/ethnicity in elementary and middle school: Trends and predictors. *Educational Researcher, 44*(6), 336-346.
- Quinn, D.M. (forthcoming). Experimental effects of 'Achievement Gap' news reporting on

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

viewers' racial stereotypes, inequality explanations, and inequality prioritization.

Educational Researcher.

Quinn, D. M., Desruisseaux, T. M., & Nkansah-Amankra, A. (2019). "Achievement Gap"

language affects teachers' issue prioritization. *Educational Researcher*, 48(7), 484-487.

Rangvid, B.S. (2018). Gender discrimination in exam grading? Double evidence from a grading

reform and a field experiment. *Working Paper Series of Danish Centre of Applied Social*

Science Retrieved from: <http://www.forskningsdatabasen.dk/en/catalog/2395268603>

Reeves, D.B. (2008). Leading to change/effective grading practices. *Educational Leadership*,

65(5), 85-87.

Rezaei, A.R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through

writing. *Assessing Writing*, 15, 18-39.

Snyder, T.D., de Brey, C., and Dillow, S.A. (2019). *Digest of Education Statistics 2017* (NCES

2018-070). National Center for Education Statistics, Institute of Education Sciences, U.S.

Department of Education. Washington, DC

Sprietsma, M. (2013). Discrimination in grading: Experimental evidence from primary school

teachers. *Empirical Economics*, 45, 523-538.

Steele, C.M. (2011). *Whistling Vivaldi: How stereotypes affect us and what we can do*. W.W.

Norton & Company.

Taylor, V. J., & Walton, G. M. (2011). Stereotype threat undermines academic learning.

Personality and Social Psychology Bulletin, 37(8), 1055-1067.

doi:10.1177/0146167211406506

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

- minority than for European American students? A meta-analysis. *Journal of Educational Psychology*, 99, 253-273.
- Terrier, C. (2016). Boys lag behind: How teachers' gender biases affect student achievement. *IZA Discussion Paper No. 10343*. Retrieved from: <http://ftp.iza.org/dp10343.pdf>
- Tierney, R. D., Simon, M., & Charland, J. (2011). Being fair: Teachers' interpretations of principles for standards-based grading. *The Educational Forum*, 75, 210–227. doi: 10.1080/00131725.2011.577669
- Uhlmann, E.L., & Cohen, G.L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, 16, 474-480.
- van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review*, 30, 1045-1058.
- Walker, T. (2016). Teachers divided over controversial 'no zero' grading policy. *neaToday*. Retrieved from: <http://neatoday.org/2016/08/04/no-zero-policy-pro-con/>
- Warikoo, N., Sinclair, S., Fei, J., & Jacoby-Senghor, D. (2016). Examining racial bias in education: A new approach. *Educational Researcher*, 45, 508–514. doi:10.3102/0013189X16683408
- Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, 21, 1-19.
- Woodcock, A., Hernandez, P.R., Estrada, M., & Schultz, P.W. (2012). The consequences of chronic stereotype threat: Domain disidentification and abandonment. *Journal of Personality and Social Psychology*, 103, 635-646.
- Xu, K., Nosek, B. and Greenwald, A.G., (2014). Psychology data from the Race Implicit

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Association Test on the Project Implicit Demo website. *Journal of Open Psychology Data*, 2(1), p.e3. DOI: <http://doi.org/10.5334/jopd.ac>

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table 1.

Descriptive Statistics by Condition with Comparisons to Nationally-representative Estimates for Teachers in 2015-2016 School Year.

	Black student author implied		White student author implied		Diff.	P	National Estimates, 2015-2016 (excludes Pre-K Teachers)
	Mean	n	Mean	n			Mean
Female	0.637	768	0.589	781	0.048	0.054	0.766
Non-binary	0.008	768	0.008	781	0.000	0.977	
<i>Teacher race</i>						0.301	
Black	0.098	768	0.090	781	0.008	0.588	0.067
White	0.685	768	0.707	781	-0.022	0.349	0.801
Asian	0.043	768	0.037	781	0.006	0.558	0.023
Latinx	0.085	768	0.063	781	0.022	0.099	0.088
Other race	0.007	768	0.003	781	0.004	0.247	0.006
Multi-racial	0.083	768	0.101	781	-0.018	0.226	0.014
<i>Teaching assignment</i>						0.014	
Pre-K	0.145	768	0.163	781	-0.018	0.324	
K-2	0.168	768	0.125	781	0.042	0.018	
Grade 3-5	0.188	768	0.150	781	0.038	0.048	0.476 (ES)
Grade 6-8	0.150	768	0.184	781	-0.035	0.068	0.178 (MS)
Grade 9-12	0.350	768	0.378	781	-0.027	0.262	0.287 (HS)
<i>School demographics</i>						0.575	
Not primarily any race/ethnicity	0.210	768	0.202	781	0.007	0.722	0.449 (Sch is <50% White)
Primarily Asian	0.016	768	0.017	781	-0.001	0.874	
Primarily Black	0.126	768	0.128	781	-0.002	0.918	
Primarily Latinx	0.120	768	0.092	781	0.028	0.078	
Primarily Native American	0.010	768	0.009	781	0.001	0.770	
Primarily White	0.518	768	0.552	781	-0.034	0.185	
<i>Years in field of edu.</i>						0.135	
< 1 year	0.039	768	0.046	781	-0.007	0.494	0.099 (<3 years)
1-3 years	0.142	768	0.133	781	0.009	0.617	
4-6 years	0.195	768	0.188	781	0.007	0.723	

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

							0.283 (3-9 years)
7-10 years	0.220	768	0.262	781	-0.042	0.051	
11-15 years	0.160	768	0.164	781	-0.004	0.842	
							0.393 (10-20 years)
16-20 years	0.109	768	0.072	781	0.038	0.010	
							0.225 (>20 years)
Over 20 years	0.134	768	0.134	781	0.000	0.985	
<i>Outcomes</i>							
	2.990 (1.233)	768	3.083 (1.280)	781	-0.094	0.143	
Grade level rating Grade level (binary at or above vs. below)	0.307 2.346 (0.769)	768	0.351 2.348 (0.757)	781	-0.044	0.068	
Rubric rating Rubric rating (binary)	0.362	768	0.373	781	-0.011	0.665	
<i>Racial attitude measures (moderators)</i>							
	0.430 (1.000)	345	0.390 (0.975)	330	0.040	0.598	
Competence IAT Attitude Thermometer (White-Black)	0.079 (1.419)	768	0.058 (1.466)	781	0.022	0.766	

Note. SDs are in parentheses. Variables with no SD entry are binary indicators for the row category (i.e., these mean values represent proportions). *p*-values are for test of the null hypothesis of no difference across conditions; *p*-values in variable rows are from t-tests, *p*-values in category rows are from chi-square tests. National estimates come from Snyder, de Brey, & Dillow (2019), tables 209.1 (race/ethnicity, gender, years in edu field) and 209.24 (grade-level, school demographics). Pre-K teachers are not included in sample for national estimates, but are included in our sample. In national estimates, 6% of teachers teach combined grade-levels. ES=elementary school teachers; MS=middle school teachers; HS=high school teachers.

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table 2.

Estimates of racial bias in evaluating student writing using general grade-level scale versus specific evaluation criteria (linear probability models).

	General Grade-level Scale			Specific Eval Criteria			
	Bias Estimate (Dashawn-Connor)		Adj Mean (Connor Group)	Bias Estimate (Dashawn-Connor)		Adj Mean (Connor Group)	n
	No controls	Controls		No controls	Controls		
Full Sample	-.044~ (.024)	-.047* (.024)	.353	-.011 (.025)	-.006 (.024)	.37	1549
<i>Bias Estimates by Subgroup</i>							
K-2 Teacher	-.096 (.064)	-.094 (.064)	.397	-.032 (.062)	-.035 (.063)	.328	227
Not K-2 Teacher	-.036 (.026)	-.041 (.026)	.347	-.004 (.027)	-.004 (.027)	.379	1322
Males	.01 (.037)	.002 (.036)	.283	.033 (.041)	.036 (.04)	.383	588
Females	-.081** (.031)	-.073* (.031)	.394	-.034 (.031)	-.03 (.031)	.361	949
White	-.074** (.029)	-.08** (.029)	.381	-.01 (.029)	-.006 (.029)	.371	1078
Black	-.032 (.074)	-.039 (.075)	.289	.018 (.079)	.016 (.078)	.33	145
Latinx	.017 (.089)	.033 (.096)	.297	-.141 (.094)	-.114 (.1)	.495	114
Asian	.057 (.119)	.007 (.14)	.303	-.046 (.124)	-.095 (.14)	.406	62
Multi-racial	.094 (.078)	.056 (.085)	.283	.03 (.081)	.024 (.081)	.332	143

Note. Heteroskedasticity-robust standard errors in parentheses.

General Grade-level Scale = 0/1 indicator for whether teacher rated the writing sample as “on grade-level” or above. Specific Eval Criteria = 0/1 indicator for whether teacher rated the writing sample as “recounts an event with some detail” or “provides a well-elaborated recount of an event,” versus “attempts to recount an event” or “fails to recount an event.” Bias estimates are the coefficient on the binary “Dashawn” indicator (vs. Connor). For full sample, estimates in “controls” column are from models that control for teacher gender, current grade-level assignment, race/ethnicity, teaching experience, and school racial demographics. Control estimates for subgroup models include all controls except for the variable that determines the subgroup. Adj. Mean = covariate-adjusted outcome mean in “Connor” writing sample group.

~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table 3.

Estimates of racial bias in evaluating student writing using general grade-level scale versus specific evaluation criteria by teachers' school racial demographics (linear probability models).

	General Grade-level Scale			Specific Eval Criteria			n
	Bias Estimate (Dashawn-Connor)		Adj Mean (Connor Group)	Bias Estimate (Dashawn-Connor)		Adj Mean (Connor Group)	
School Racial Demographics	No controls	Controls		No controls	Controls		
Not primarily any single race/ethnicity	-.119* (.052)	-.133* (.052)	.393	-.001 (.054)	-.022 (.054)	.378	319
Primarily Black/African American	.019 (.065)	.011 (.065)	.284	.001 (.069)	.027 (.07)	.357	197
Primarily Latinx	.011 (.069)	-.007 (.07)	.26	.004 (.075)	-.004 (.072)	.338	164
Primarily White/European American	-.037 (.033)	-.047 (.033)	.376	-.016 (.034)	-.014 (.033)	.379	829

Note. Heteroskedasticity-robust standard errors in parentheses.

General Grade-level Scale = 0/1 indicator for whether teacher rated the writing sample as “on grade-level” or above. Specific Eval Criteria = 0/1 indicator for whether teacher rated the writing sample as “recounts an event with some detail” or “provides a well-elaborated recount of an event,” versus “attempts to recount an event” or “fails to recount an event.” Bias estimates are the coefficient on the binary “Dashawn” indicator (vs. Connor). Estimates in “controls” are from models that control for teacher gender, current grade-level assignment, race/ethnicity, and teaching experience. Adj. Mean = covariate-adjusted outcome mean in “Connor” writing sample group.

~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table 4.

Linear probability models estimating interactions between measures of teachers' racial attitudes and student author's implied race (predicting teachers' evaluations of student writing).

	(1) Grade-level rating	(2) Grade-level rating	(3) Rubric rating	(4) Rubric rating
Black author	-0.0637~ (0.0375)	-0.0466~ (0.0239)	-0.0318 (0.0399)	-0.00605 (0.0245)
IAT	0.0158 (0.0280)		-0.0102 (0.0265)	
Black author*IAT	-0.0265 (0.0361)		0.0227 (0.0371)	
Racial attitude thermometer		0.0128 (0.0114)		-0.00307 (0.0115)
Black author* thermometer		-0.00685 (0.0157)		0.00464 (0.0165)
<i>N</i>	675	1549	675	1549
<i>R</i> ²	0.044	0.020	0.041	0.024

Note. Heteroskedasticity-robust standard errors in parentheses. Grade-level = 0/1 indicator for whether teacher rated the writing sample as “on grade-level” or above. Rubric = 0/1 indicator for whether teacher rated the writing sample as “recounts an event with some detail” or “provides a well-elaborated recount of an event,” versus “attempts to recount an event” or “fails to recount an event.” Models control for teacher gender, current grade-level assignment, race/ethnicity, teaching experience, and school racial demographics.

~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Appendix A. Additional Sample Comparisons

Table A1.
Descriptive Statistics Comparing Teachers Included and Excluded from the IAT Moderation Analysis.

	IAT Sample			Not in IAT Sample			Difference	<i>p</i>
	Mean	SD	n	Mean	SD	n		
Female	0.621		675	0.606		874	0.014	0.566
Non-binary	0.010		675	0.006		874	0.005	0.301
<i>Teacher race</i>								0.002
Black	0.077		675	0.106		874	-0.029	0.049
White	0.739		675	0.662		874	0.077	0.001
Asian	0.022		675	0.054		874	-0.032	0.002
Latinx	0.068		675	0.078		874	-0.010	0.471
Other race	0.001		675	0.007		874	-0.005	0.117
Multi-racial	0.092		675	0.093		874	-0.001	0.956
<i>Teaching assignment</i>								0.017
Pre-K	0.120		675	0.180		874	-0.060	0.001
K-2	0.156		675	0.140		874	0.016	0.379
Grade 3-5	0.163		675	0.173		874	-0.010	0.609
Grade 6-8	0.182		675	0.156		874	0.027	0.164
Grade 9-12	0.379		675	0.352		874	0.027	0.276
<i>School demographics</i>								0.020
Not primarily any race/ethnicity	0.185		675	0.222		874	-0.037	0.076
Primarily Asian	0.013		675	0.018		874	-0.005	0.441
Primarily Black	0.124		675	0.129		874	-0.005	0.777
Primarily Latinx	0.092		675	0.117		874	-0.025	0.115
Primarily Native American	0.004		675	0.014		874	-0.009	0.064
Primarily White	0.581		675	0.500		874	0.081	0.002
<i>Years in field of edu.</i>								<0.001
< 1 year	0.021		675	0.059		874	-0.039	<0.001
1-3 years	0.144		675	0.133		874	0.011	0.534
4-6 years	0.181		675	0.200		874	-0.019	0.334
7-10 years	0.215		675	0.262		874	-0.047	0.031
11-15 years	0.206		675	0.128		874	0.078	<0.001
16-20 years	0.096		675	0.086		874	0.010	0.476
Over 20 years	0.138		675	0.132		874	0.006	0.723
<i>Outcomes</i>								
Grade level rating	2.921	1.126	675	3.126	1.344	874	-0.204	0.001

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Grade level (binary at or above vs. below)	0.287	0.453	675	0.362	0.481	874	-0.074	0.002
Rubric rating	2.361	0.714	675	2.336	0.799	874	0.025	0.521
Rubric rating (binary)	0.344	0.475	675	0.386	0.487	874	-0.042	0.090

Note. Variables with no SD entry are binary indicators. *p*-values are for test of the null hypothesis of no difference across conditions; entries in variable rows are from t-tests, entries in category rows are from chi-square tests.

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table A2.

Descriptive Statistics by Experimental Condition among Teachers Included in the IAT Moderation Analysis.

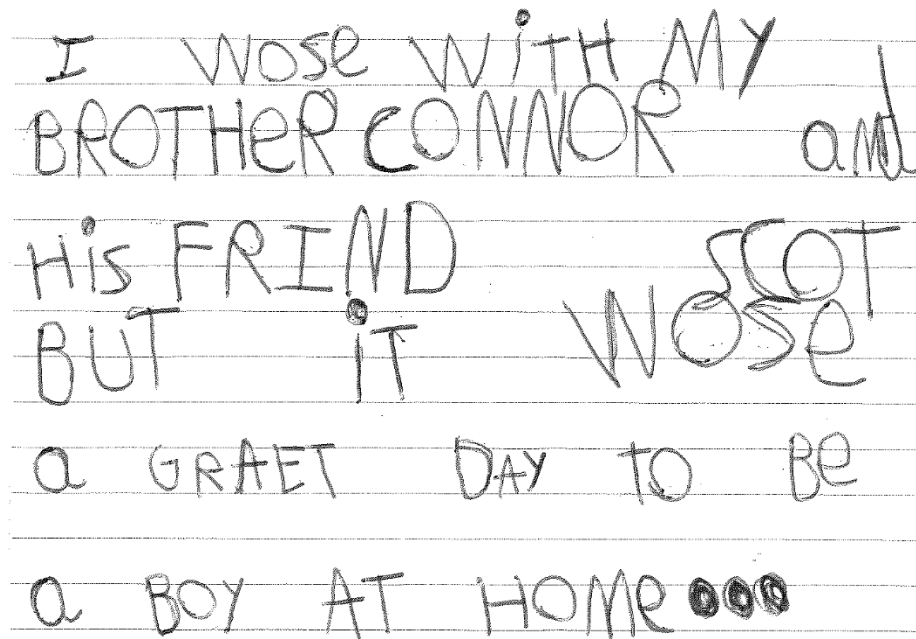
	Dashawn			Connor			Differenc	
	Mean	SD	n	Mean	SD	n	e	p
Female	0.658		345	0.582		330	0.076	0.042
Non-binary	0.009		345	0.012		330	-0.003	0.661
<i>Teacher race</i>								0.912
Black	0.075		345	0.079		330	-0.003	0.868
White	0.739		345	0.739		330	0.000	0.994
Asian	0.026		345	0.018		330	0.008	0.487
Latinx	0.067		345	0.070		330	-0.003	0.876
Other race	0.003		345	0.000		330	0.003	0.328
Multi-racial	0.090		345	0.094		330	-0.004	0.855
<i>Teaching assignment</i>								0.065
Pre-K	0.099		345	0.142		330	-0.044	0.080
K-2	0.186		345	0.124		330	0.061	0.028
Grade 3-5	0.180		345	0.145		330	0.034	0.229
Grade 6-8	0.171		345	0.194		330	-0.023	0.441
Grade 9-12	0.365		345	0.394		330	-0.029	0.443
<i>School demographics</i>								0.901
Not primarily								
any race/ethnicity	0.191		345	0.179		330	0.013	0.676
Primarily Asian	0.009		345	0.018		330	-0.009	0.283
Primarily Black	0.125		345	0.124		330	0.000	0.988
Primarily Latinx	0.090		345	0.094		330	-0.004	0.855
Primarily Native								
American	0.006		345	0.003		330	0.003	0.590
Primarily White	0.580		345	0.582		330	-0.002	0.956
<i>Years in field of edu.</i>								0.183
< 1 year	0.032		345	0.009		330	0.023	0.038
1-3 years	0.139		345	0.148		330	-0.009	0.730
4-6 years	0.188		345	0.173		330	0.016	0.597
7-10 years	0.194		345	0.236		330	-0.042	0.183
11-15 years	0.191		345	0.221		330	-0.030	0.338
16-20 years	0.113		345	0.079		330	0.034	0.132
Over 20 years	0.142		345	0.133		330	0.009	0.744
<i>Outcomes</i>								
Grade level								
rating	2.858	1.110	345	2.988	1.141	330	-0.130	0.134

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Grade level (binary at or above vs. below)	0.252	0.435	345	0.324	0.469	330	-0.072	0.039
Rubric rating	2.336	0.709	345	2.388	0.720	330	-0.052	0.348
Rubric rating (binary)	0.325	0.469	345	0.364	0.482	330	-0.039	0.287
<i>Interaction variables</i>								
Competence								
IAT	0.430	1.000	345	0.390	0.975	330	0.040	0.598
Attitude								
Thermometer (White-Black)	0.113	1.228	345	0.124	1.353	330	-0.011	0.910

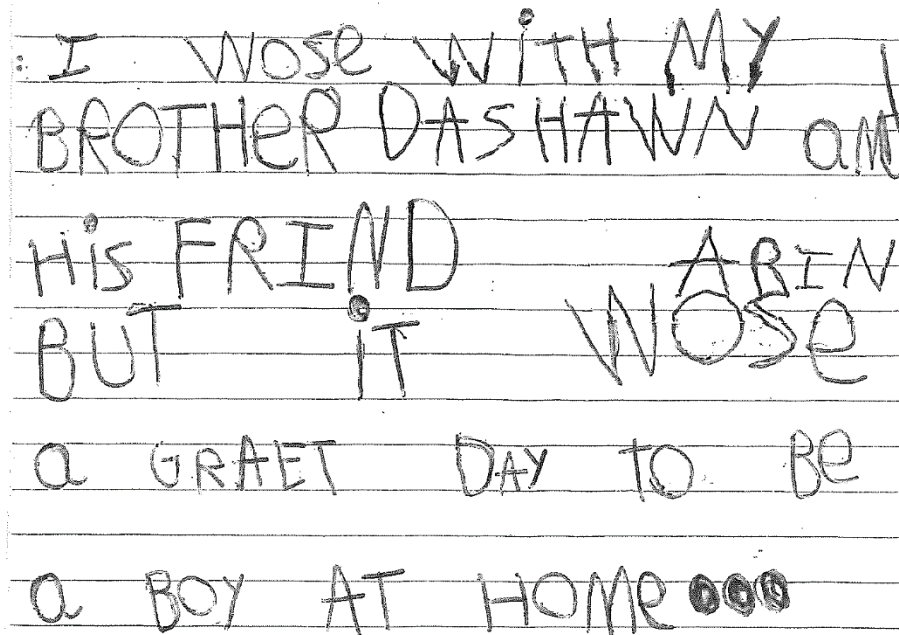
Note. Variables with no SD entry are binary indicators. *p*-values are for test of the null hypothesis of no difference across conditions; entries in variable rows are from t-tests, entries in category rows are from chi-square tests.

Appendix B. Experimental Materials.



I WOSE WITH MY
BROTHER CONNOR and
HIS FRIEND
BUT IT WOSE SCOT
A GRAET DAY TO BE
A BOY AT HOME

Figure B1. Student writing sample with Anglo name. Sample purportedly written by student in fall of second grade in response to the prompt to write about their weekend.



I WOSE WITH MY
BROTHER DASHAWN and
HIS FRIEND
BUT IT WOSE ARIN
A GRAET DAY TO BE
A BOY AT HOME

Figure B2. Student writing sample with distinctively Black name. Sample purportedly written by student in fall of second grade in response to the prompt to write about their weekend.

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Appendix C. Ordered Logistic Regression Models with 7-point Grade-level Scale and 4-point Rubric

Table C1. Ordered logistic regression models estimating grading bias and moderation by measures of teachers' racial attitudes.

	(1) Grade-level rating	(2) Grade-level rating	(3) Grade-level rating
Black author	-0.182* (0.0916)	-0.199 (0.151)	-0.155~ (0.0915)
Female	0.459*** (0.0994)		
IAT		0.0706 (0.104)	
Black author X IAT		-0.0952 (0.138)	
Racial attitude thermometer			0.0167 (0.0521)
Black author X thermometer			-0.0202 (0.0758)
cut1 Constant	-1.988*** (0.117)	-2.166*** (0.153)	-2.241*** (0.0994)
cut2 Constant	-0.482*** (0.0949)	-0.772*** (0.120)	-0.753*** (0.0719)
cut3 Constant	0.924*** (0.0974)	0.818*** (0.119)	0.636*** (0.0703)
cut4			

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Constant	2.546*** (0.126)	2.628*** (0.179)	2.251*** (0.0986)
cut5 Constant	3.414*** (0.159)	4.015*** (0.316)	3.121*** (0.137)
cut6 Constant	4.110*** (0.206)	5.037*** (0.516)	3.817*** (0.188)
<i>N</i>	1549	675	1549

Note. Heteroskedasticity-robust standard errors in parentheses

~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table C2. Ordered logistic regression models estimating grading bias and moderation by measures of teachers' racial attitudes.

	(1) Rubric	(2) Rubric	(3) Rubric
Black author	-0.00883 (0.0976)	-0.150 (0.166)	-0.00662 (0.0980)
Female	-0.205* (0.103)	-0.623*** (0.166)	-0.203* (0.103)
IAT		-0.113 (0.112)	
Black author X IAT		0.121 (0.154)	
Racial attitude thermometer			0.00830 (0.0515)
Black author X thermometer			-0.0332 (0.0824)
<hr/>			
cut1			
Constant	-2.347*** (0.121)	-3.269*** (0.230)	-2.345*** (0.121)
<hr/>			
cut2			
Constant	0.412*** (0.0942)	0.164 (0.156)	0.414*** (0.0944)
<hr/>			
cut3			
Constant	2.332*** (0.123)	2.014*** (0.198)	2.334*** (0.123)
<hr/>			
<i>N</i>	1549	675	1549

Note. Heteroskedasticity-robust standard errors in parentheses

~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Appendix D. Competence IAT: Description and Development

Implicit stereotypes. To measure implicit racial stereotypes, I developed an implicit association test relating race (Black/White) and competence. In its original form, the traditional Black/White IAT is a valence measure, meaning that it measures the relative strength of one's pairing of a positive versus negative valence with White people versus with Black people. It does this through a computerized timed classification task that compares how quickly and accurately test-takers can classify stimuli representing European Americans (e.g., photographs of faces) when the race category is paired with a good vs. bad valence term (e.g., "joy" vs. "hurt") to how quickly and accurately they can classify stimuli representing African Americans.

For my competence IAT, I use the categories "African American" and "European American." Following Fiske et al. (2002), I use the categories "Competent" and "Incompetent," and the competence target words "intelligent," "confident," "capable," and "efficient." I use the incompetence target words "disorganized," "unqualified," "stupid," and "unskilled" (inspired by Vitriol, Ksiazkiewicz, & Farhart [2018]). The stimuli included photographs of Black and White adolescents (4 male, 4 female for each racial group), obtained through Getty images and piloted on Amazon's MTurk platform to ensure that the age and race of the photographed subjects were perceived as intended. Using a selection of photographs in which the perceived race was as intended and in which subjects' perceived ages were similar across races/genders, I built the competence IAT using the iatgen online software (Carpenter et al., 2018).

Prior to this study, I tested whether the competence IAT differed from the traditional Black-White valence IAT by conducting a pilot on MTurk in which respondents (target sample of $n=300$; 40 dropped for excessive speed, yielding final $n=260$) were randomly assigned to complete my competence IAT or the traditional Black/White IAT (respondents were paid \$1.00).

In the pilot, internal consistency (based on split-half with Spearman-Brown correction) was .86 for the competence IAT and .85 for the traditional IAT (error rates were .086 and .091 for competence and traditional, respectively). A t-test showed that scores on the traditional and competence IAT were significantly different ($p=.013$), suggesting that the competence IAT is measuring a unique dimension of implicit stereotyping compared to the traditional race IAT. On both tests, the average respondent showed significant pro-White bias, though the magnitude was smaller for the competence IAT (average traditional d -score = .42; average competence d -score = .30).

Validity of Black/White competence IAT. In a separate study (Quinn, forthcoming), I collected validity evidence for my implicit measure by correlating respondents' IAT scores with a variety of measures (described below) and fitting a series of regression models predicting individuals' IAT d -scores (divided by the sample SD). To establish known-groups validity, I fit a set of models with including indicators for respondent race/ethnicity.

Prioritization of educational inequality. Following Valant & Newark (2016), I measured the extent to which respondents prioritized racial achievement disparities with the item, "As you may know, there is a racial academic achievement gap between Black and White students in the US. Thinking about all of the important issues facing the country today, how much of a priority do you think it is to close the racial academic achievement gap between Black and White students?" Answer choices were on a 5-point scale (1=not a priority; 2=low priority; 3=medium priority; 4=high priority; 5=essential). I expected this item to negatively correlate with the IAT.

Explanations of educational inequality. I surveyed respondents on their beliefs about the sources of racial achievement disparities with the item, "To what extent do you believe each

of these factors is responsible for the racial academic achievement gap between Black and White students?" Respondents were asked to rate the contributions of the following possible explanations (with order randomized): School quality, student motivation, parenting, discrimination and racism, genetics, neighborhood environments, home environments, and income levels. Answer choices were on a 5-point scale (1= not at all; 2= slightly; 3=somewhat; 4=quite; 5=extremely; items were inspired by Valant & Newark [2016]).

I created two indices from the explanation items using principal components analysis (PCA). The PCA revealed two components with eigenvalues above 1. The first component (eigenvalue = 3.69) positively weighted all items and explained 46% of the total variation. The second component – which I call “non-structural” (eigenvalue = 1.21, explaining 15% of total variation) – positively weighted the motivation, parenting, genetics, and home environment explanations and negatively weighted the school quality, discrimination, and income explanations (with a negative, but near-zero, weight for neighborhoods). As such, people who scored highly on this index tended to discount structural explanations for racial achievement disparities, in favor of cultural and genetic explanations. I expected the non-structural items to negatively correlate with the IAT and structural items to positively correlate with the IAT.

Stereotyping. To measure the extent to which respondents explicitly stereotyped Black Americans as lacking formal education, I administered the following item: “The national high school graduation rate for White students is 86%. What is your best guess of what the national high school graduation rate is for Black students? Type the percentage in the box below.” (actual Black graduation rate is approximately 78% [Murnane, 2013]). I expected this item to negatively correlate with the IAT.

In Table D1, I present the correlations of the competence IAT scores with these validation items. Demonstrating initial validity evidence, respondents' competence IAT d-scores were significantly correlated with several survey measures as expected. Magnitudes were small but similar to the average of $r=.12$ found in a meta-analysis of implicit and explicit racial attitudes (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). Respondents with more pro-White implicit competence bias showed lower guesses for Black students' high school graduation rates ($r = -.094$), gave less priority to closing racial achievement gaps ($r = -.185$), were less likely to believe that school quality played a larger role in racial achievement inequality ($r = -.175$), and were less likely to believe discrimination and racism played an important role in racial achievement inequality ($r = -.154$). People showing more pro-White competence bias on the IAT may also be more likely to believe that parenting plays an important role in racial achievement disparities ($r = .085, p < .10$) and may be less likely to believe that income plays an important role ($r = -.08, p < .10$). Implicit bias did not predict the extent to which respondents believed that motivation, genetics, neighborhood, or home environment helped explain racial achievement disparities.

In Table D2, I present additional validity evidence through OLS regression models. This table provides known-groups validity evidence by demonstrating that, unlike White respondents, Black respondents have implicit associations of Black students as being more competent than White students.

Table D1.

Correlations of IAT competence d-scores with other outcome variables.

	IAT d-score
Graduation rate guess	-0.09*
Gap Priority	-0.19***
Explanation index	-0.06
Non-structural explanation index	0.21***
Explanation: school quality	-0.18***
Explanation: student motivation	-0.01
Explanation: parenting	0.08~
Explanation: discrimination & racism	-0.15***
Explanation: genetics	0.07
Explanation: neighborhood environment	-0.05
Explanation: home environment	0.01
Explanation: income	-0.08~

~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note. Graduation rate guess = guess of Black HS graduation rate; Gap priority = how much of a priority believes is to close academic achievement gap between Black and White students (1=not a priority to 5=essential). Explanation index = PCA index positively weighting all explanations for achievement gaps; Non-structural explanation index = PCA index positively weighting non-structural explanations for gap; Gap explanations items give extent to which respondent believes that factor is responsible for racial academic achievement gap between Black and White students (1=not at all; 2=slightly; 3=somewhat; 4=quite; 5=extremely)

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table D2.
OLS Regression Models Predicting Implicit Competence Stereotypes

	(1)	(2)	(3)	(4)	(5)
	D-score (std)	D-score (std)	D-score (std)	D-score (std)	D-score (std)
Educator	-0.217 (0.134)			-0.208 (0.133)	-0.201 (0.132)
Black	-1.034*** (0.113)			-0.972*** (0.113)	-0.982*** (0.113)
Latinx	-0.313~ (0.187)			-0.379* (0.184)	-0.369* (0.183)
Asian	-0.185 (0.175)			-0.211 (0.173)	-0.200 (0.172)
Other Race	-0.831 (0.653)			-0.791 (0.645)	-0.827 (0.641)
Multi-racial	-0.664*** (0.157)			-0.661*** (0.156)	-0.640*** (0.155)
American Indian	-0.984* (0.461)			-0.764~ (0.458)	-0.849~ (0.452)
Female	0.0186 (0.0899)			0.0877 (0.0900)	0.0704 (0.0886)
Non-binary	-0.894* (0.420)			-0.696~ (0.420)	-0.748~ (0.414)
Grad Guess		-0.00329 (0.00202)	-0.00356~ (0.00201)	-0.00430* (0.00189)	-0.00463* (0.00188)
Gap Priority (std)		-0.105* (0.0514)	-0.120* (0.0502)	-0.113* (0.0477)	-0.123** (0.0467)
Sch. Quality		-0.139* (0.0553)		-0.113* (0.0519)	
Motivation		-0.0232 (0.0491)		-0.0118 (0.0455)	
Parenting		0.115* (0.0569)		0.114* (0.0528)	
Discrimination		-0.0754		-0.0237	

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

		(0.0477)		(0.0452)	
Genetics		0.0444		0.0275	
		(0.0333)		(0.0315)	
Neighborhood d enviro		-0.0203		-0.0245	
		(0.0581)		(0.0542)	
Home enviro		0.0219		-0.000807	
		(0.0681)		(0.0637)	
Income		0.0358		0.0497	
		(0.0511)		(0.0479)	
Explanation index (std)			-0.0239		0.0199
			(0.0491)		(0.0459)
Non- structural index (std)			0.164***		0.103*
			(0.0452)		(0.0427)
Constant	0.613***	0.738*	0.564***	0.768**	0.838***
	(0.0931)	(0.291)	(0.129)	(0.278)	(0.141)
<i>N</i>	514	514	514	514	514
<i>R</i> ²	0.181	0.088	0.074	0.234	0.223
<i>F</i>	11.09	4.396	8.139	7.541	10.21

Standard errors in parentheses. Outcome is d-score divided by its SD. Data collected in validation sample (Quinn, forthcoming) in which respondents were also randomly assigned to a treatment of viewing brief education-related clips on YouTube. Models also controls for random assignment to treatment condition.

~ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Appendix E. Logistic Regression Models

Table E1. Logistic regression models estimating main effects of student author's implied race on teachers' evaluations.

	(1) Grade-level rating	(2) Rubric rating
Black author	-0.214* (0.109)	-0.0344 (0.106)
Female	0.331** (0.113)	-0.238* (0.108)
Constant	-0.814*** (0.102)	-0.383*** (0.0966)
<i>N</i>	1549	1549

Standard errors in parentheses

~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table E2. Logistic regression models estimating main effects of student author's implied race on teachers' grade-level evaluations, by teacher race/ethnicity and gender.

	(1) Male	(2) Female	(3) White	(4) Black	(5) Latinx	(6) Asian	(7) Multi-racial
Black author	0.0492 (0.183)	-0.353** (0.136)	-0.347** (0.130)	-0.253 (0.383)	0.0831 (0.413)	0.317 (0.562)	0.429 (0.365)
Female			0.326* (0.135)	0.839* (0.401)	-0.0306 (0.418)	0.388 (0.590)	0.177 (0.372)
Constant	-0.948*** (0.126)	-0.415*** (0.0953)	-0.696*** (0.122)	-1.374*** (0.359)	-0.802* (0.381)	-1.240* (0.598)	-1.116*** (0.333)
<i>N</i>	588	949	1078	145	114	62	143

Standard errors in parentheses

~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Table E3. Logistic regression models estimating main effects of student author's implied race on teachers' rubric evaluations, by teacher race/ethnicity and gender.

	(1) Male	(2) Female	(3) White	(4) Black	(5) Latinx	(6) Asian	(7) Multi-racial
Black author	0.139 (0.169)	-0.149 (0.137)	-0.0293 (0.127)	0.0447 (0.355)	-0.565 (0.388)	-0.198 (0.536)	0.152 (0.355)
Female			-0.340** (0.130)	0.369 (0.359)	-0.0803 (0.396)	0.0248 (0.555)	-0.318 (0.357)
Constant	-0.472*** (0.116)	-0.562*** (0.0970)	-0.316** (0.116)	-0.905** (0.319)	0.0834 (0.355)	-0.510 (0.542)	-0.539~ (0.306)
<i>N</i>	588	949	1078	145	114	62	143

Standard errors in parentheses ~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table E4. Logistic regression models estimating main effects of student author's implied race on teachers' grade-level evaluations, by demographics of teacher's school.

	(1) Not prim. any race/ethnicity	(2) Primarily Black	(3) Primarily Latinx	(4) Primarily White
Black author	-0.581* (0.247)	0.0989 (0.315)	0.0653 (0.364)	-0.180 (0.146)
Female	0.993*** (0.280)	-0.0828 (0.322)	-0.0653 (0.377)	0.330* (0.150)
Constant	-1.135*** (0.260)	-0.898** (0.287)	-1.061** (0.348)	-0.717*** (0.133)
<i>N</i>	319	197	164	829

Standard errors in parentheses

~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Table E5. Logistic regression models estimating main effects of student author's implied race on teachers' rubric evaluations, by demographics of teacher's school.

	(1) Not prim. any race/ethnicity	(2) Primarily Black	(3) Primarily Latinx	(4) Primarily White
Black author	-0.00402 (0.232)	0.0379 (0.298)	-0.000334 (0.336)	-0.0501 (0.145)
Female	0.102 (0.245)	-0.409 (0.302)	0.136 (0.353)	-0.377** (0.146)
Constant	-0.611** (0.230)	-0.304 (0.265)	-0.773* (0.326)	-0.279* (0.127)
<i>N</i>	319	197	164	829

Standard errors in parentheses

~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Table E6. Logistic regression models estimating interactions of measures of teachers' racial attitudes with student author's implied race predicting teachers' evaluations.

	(1) Grade-level rating	(2) Grade-level rating	(3) Rubric rating	(4) Rubric rating
Black author	-0.362~ (0.189)	-0.212~ (0.109)	-0.162 (0.178)	-0.0354 (0.106)
IAT	0.102 (0.122)		-0.0741 (0.119)	
Black author X IAT	-0.113 (0.174)		0.0883 (0.167)	
Female	0.665*** (0.187)	0.329** (0.113)	-0.623*** (0.166)	-0.238* (0.108)
Racial attitude thermometer		0.0396 (0.0516)		-0.0148 (0.0506)
Black author X thermometer		-0.0318 (0.0760)		0.0180 (0.0732)
Constant	-1.183*** (0.176)	-0.816*** (0.103)	-0.182 (0.153)	-0.382*** (0.0967)
<i>N</i>	675	1549	675	1549

Standard errors in parentheses

~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Appendix F. Bias on Relative Grade-level Scale for Gender-by-race Subgroups.

Table F1. Bias estimates from relative grade-level scale by teacher gender and race.

Dashawn-Connor	Connor	n	Teacher Race
Male Teachers			
-.037 (.046)	0.317	404	White
.131 (.119)	0.229	60	Multi-racial
.211 (.14)	0.217	44	Latino
.063 (.196)	0.222	23	Asian
.029 (.098)	0.171	65	Black
Female Teachers			
-.102** (.037)	0.419	674	White
.064 (.104)	0.295	83	Multi-racial
-.112 (.116)	0.385	70	Latina
.068 (.155)	0.3	39	Asian
-.111 (.107)	0.4	80	Black

Note. Estimates from linear probability models predicting whether teacher rated writing as on grade-level or above (1) versus below (0). “Connor” column is proportion rating “Connor” sample as on grade-level or above; “Dashawn-Connor” is Dashawn-Connor difference. Standard errors in parentheses.

~ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

EXPERIMENTAL EVIDENCE ON TEACHERS' RACIAL BIAS

Additional References (Appendices)

- Carpenter, T., Pogacar, R., Pullig, C., Kouril, M., LaBouff, J., Aguilar, S. J., ...
Chakroff, A. (2018, April 3). Conducting IAT Research within Online Surveys: A
Procedure, Validation, and Open Source Tool. <https://doi.org/10.31234/osf.io/6xdyj>
- Fiske, S.T., Cuddy, A.J.C., Glick, P., & Xu, J. (2002) A model of (often mixed) stereotype
content: Competence and warmth respectively flow from perceived status and
competition. *Journal of Personality and Social Psychology*, 82, 878-902.
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009).
Understanding and using the Implicit Association Test: III. Meta-analysis of predictive
validity. *Journal of personality and social psychology*, 97(1), 17.
- Murnane, R.J. (2013). U.S. high school graduation rates: Patterns and explanations. *Journal of
Economic Literature*, 51, 370-422.
- Quinn, D.M. (forthcoming). Experimental effects of 'Achievement Gap' news reporting on
viewers' racial stereotypes, inequality explanations, and inequality prioritization.
Educational Researcher.
- Valant, J., & Newark, D. A. (2016). The politics of achievement gaps: US public
opinion on race-based and wealth-based differences in test scores. *Educational
Researcher*, 45(6), 331-346.
- Vitriol, J.A., Ksiazkiewicz, A., & Farhart, C.E. (2018). Implicit candidate traits in the 2016
U.S. presidential election: Replicating a dual-process model of candidate evaluations.
Electoral Studies, 54, 261-268.