How to Reduce Racial Bias in Grading

New research supports a simple, low-cost teaching tool

SCHOOLS AND POLICYMAKERS are mandating new anti-bias training for teachers in an attempt to improve racial attitudes. Decades of research have shown that teachers often give racially biased evaluations of student work and that biased evaluations can affect students' future learning and course-taking decisions. However, less is known about what school leaders can do to correct this problem. Research does not show current forms of anti-bias training to be especially promising in changing behavior.

There is, though, a relatively straightforward, if often overlooked, way to diminish the impact of teachers' racial biases in student evaluation: standardizing grading rubrics. To gauge the potential impact of a standardized rubric on grading bias, I conducted an experiment comparing how teachers graded two identical second-grade writing samples: one presented as the work of a Black student, and one as the work of a white student.

My experiment found that teachers gave the white student better marks across the board—with one exception. When teachers used a grading rubric with specific criteria, racial bias all but disappeared. When teachers evaluated student writing using a general grade-level scale, they were 4.7 percentage points more likely to consider the white child's writing at or above grade level compared to the identical writing from a Black child. However, when teachers used a grading rubric with specific criteria, the grades were essentially the same.

The experiment also included a series of questions asking teachers about their background and their racial attitudes. In exploratory analyses examining bias by teachers' own race, gender, and the racial makeup of the schools where they teach, I found larger bias in grading by white and female teachers, who were less likely to rate the Black child's writing as being on grade level compared to the white child's writing. However, I didn't find any connection between my measures of teachers' implicit and explicit racial attitudes and the differences in grading the Black and white student writing samples.

This experiment suggests that racial stereotypes can influence the scores teachers assign to student work. But stereotypes seem to have less influence on teachers' evaluations when specific grading criteria are established in advance. New instructional practices and tools, such as standards-based grading rubrics and mastery-based grading with specific criteria, present potentially effective approaches to promoting racial equity in schools.

by DAVID M. QUINN

JEAN-FRANCOIS PODEVIN

Limiting opportunities for biased decisions may have more immediate impact on equitable student evaluation than current forms of anti-bias training.

Building a Grading-Bias Experiment

My experiment took the form of a web-based survey, including demographic questions, a two-part grading task, and a test to measure racial attitudes. I contracted with a private survey provider to recruit a multi-state sample of U.S. schoolteachers. Some 1,799 unique users responded to a survey invitation. Of those, 1,549 teachers working in preschool through 12th grade completed the main survey tasks and were compensated directly by the company for participating. Their responses form the basis of my analysis.

At the start of the survey, teachers were informed that the researcher was interested in learning how educators evaluate student writing. 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. Overt statements of a student's race in a grading experiment could arouse suspicion among research participants and affect their responses.

Respondents also answered questions about their gender, race, and number of years in the field, as well as the grade that they teach and the racial composition of their school. Overall, 69 percent were white and 54 percent taught in a predominantly white school. By comparison, about 79 percent of U.S. teachers are white and approximately 45 percent of all U.S. teachers work in schools that are less than 50 percent white, according to federal data from 2017.

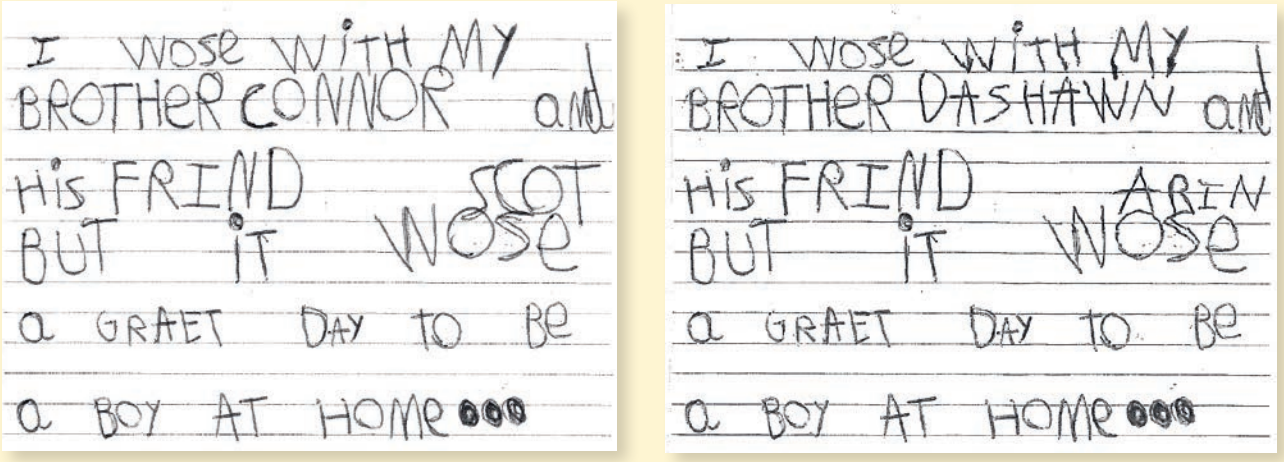
Two Grading Tasks

Teachers were randomly assigned to receive one of two versions of a writing sample. The writing sample was presented in a child's handwriting and was purportedly by a male student in the fall of second grade. It took the form of a brief personal narrative in response to a prompt to write about his weekend and mentioned spending time with his brother and a friend. The versions were identical in all but one aspect: each used different names for the brother to signal either a Black or a white student author (see Figure 1). The name choices came from a list of the most racially distinct names reported by Steven Levitt and Stephen J. Dubner in *Freakonomics*. In one version, the student author refers to his brother as "Dashawn," signaling a Black author; in the other, his brother is called "Connor," signaling a white author.

Teachers were first asked to rate the writing sample on a relative grade-level scale with seven options, from weak to strong performance: far below grade level, below grade level, and slightly below grade level; at grade level; or slightly above

One Writing Sample, Two Student Races (Figure 1)

In an experiment, teachers were asked to assess one of the two writing samples below, which were presented as the work of a second-grade boy asked to write about his weekend. The work is identical except for the names mentioned—either "Dashawn," suggesting it was written by a Black student, or "Connor," suggesting it was written by a white student.



SOURCE: Author

Teachers were 4.7 percentage points more likely to consider the white child's writing at or above "grade level" compared to the identical writing from a Black child.

grade level, above grade level, and far above grade level. Performance criteria were not explicitly defined.

Then, they were asked to rate the writing sample again. This time, teachers were given a rubric with more clearly defined performance criteria for a personal narrative. The rubric included four possible ratings, from weak to strong: fails to recount an event, attempts to recount an event, recounts an event with some detail, or provides a well-elaborated recount of an event. The rubric appeared after the grade-level scale, on a separate screen and without the option to return to the earlier screen. This was designed to ensure that teachers' ratings on the grade-level scale were not influenced by the rubric's criteria.

Substantively, these evaluation measures differ in two important respects. The grade-level scale is general in the sense that it does not specify what dimensions the rater should consider, such as grammar, spelling, creativity, or organization. It also does not clearly specify the gradations among the seven possible ratings, or how a teacher should determine whether the writing is "slightly above grade level" versus "above grade level." In contrast, the rubric specifies which domain teachers should evaluate—in this case, how well the writer recounts an event—and provides more specific descriptors to guide teachers in their rating choices along a four-point scale.

Assessing Racial Attitudes

After the grading exercise, the survey then attempted to collect data on respondents' racial attitudes. This presented a challenge. On one hand, if questions or activities designed to reveal racial attitudes were administered before teachers see the writing sample, the act of completing the racial attitude measures could influence their grading. In particular, the experiment could produce "demand effects," in which teachers adjust their ratings to match what they view as a socially desirable response, such as being particularly careful to show no racial bias. However, if respondents complete the racial attitude measures after viewing the writing sample, the writing sample may influence their racial attitude scores. I opted for the second option, viewing this as less damaging to the experiment overall.

To measure teachers' implicit stereotypes of white and Black students, I adapted a traditional implicit association test to assess respondents' associations between race and competence. These computer-based tests ask respondents to react quickly to ideas and images by assigning them to one of two categories, such as "good" and "bad." In my test, participants identified photos of students' faces as either "African American" or "European American" by pressing a right- or left-hand key on a computer keyboard. Then, those same keys were also used to assign words like intelligent, confident, disorganized, and unskilled to one of two categories: either "Competent" or "Incompetent." The test combines the racial categories and competency categories in various combinations, with the right-hand key used for "European American" and "Incompetent" on one round, and for "African American" and "Incompetent" on the next round, for example.

Implicit association test scores are calculated by comparing response speeds by category. In this case, the relevant question is how long it takes a respondent to assign descriptive "Competence" or "Incompetence" words when those categories use the same keyboard stroke as "African American" or "European American." An implicit association result would indicate a preference for African Americans over European Americans, for example, if a participant's responses were faster when "African American" and "Competence" were assigned to the same key. In my study, just 675 teachers completed the full test and produced valid scores that are included in my analysis. The other half either abandoned the survey or responded to test items so quickly that valid scores could not be calculated.

Finally, teachers were asked to respond to traditional "feeling thermometer" questions in which they rated their feelings toward African Americans and 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 in the warmth of each individual teacher's feelings about white and Black Americans. A positive score indicates a preference for white Americans and a negative score indicates a preference for Black Americans. Some 1,549 teachers completed the feeling thermometer questions.

Results by Grading Tool

To compare teachers' grades using both grading tools, I sort grades into two groups based on whether their rating is above or below a cut-off point on each scale. On the vague grade-level scale, I consider how many teachers rate each sample as "at grade level" or above. On the rubric, I look at how many teachers rate the sample as "recounts an event

with some detail" or better. However, results were robust to a variety of other analytic choices.

Teachers shown the "Dashawn" version of the writing sample are 4.7 percentage points less likely to rate it as being on grade-level or above compared to teachers shown the "Connor" version (see Figure 2). Some 35 percent of respondents rate the version written by a white student at grade-level or above compared with about 30 percent for the version written by a Black student. However, when those same teachers use a rubric with specific grading criteria, they give essentially identical ratings to the Black and white authors—about 37 percent rate both the "Dashawn" and "Connor" versions as "recounts an event with some detail" or better.

In exploratory analyses, I also investigate differences in grading based on the gender and race of the teacher. Prior research has found that teachers show preference for students with identities similar to their own. In particular, white teachers tend to have lower expectations for Black students than for similar white students (see "The Power of Teacher Expectations," *research,* Winter 2018). I see evidence of this in my experiment as well, though only when teachers apply the vague grade-level scale. In all groups, when teachers use a specific grading rubric, estimates of bias are small and not significant.

Female teachers assessing a young boy's writing sample show racial bias in their grading, but male teachers do not. Whereas females are 7 percentage points less likely to rate the "Dashawn" sample as being on grade level than the "Connor" sample, the difference for male teachers is small and not statistically significant (see Figure 3).

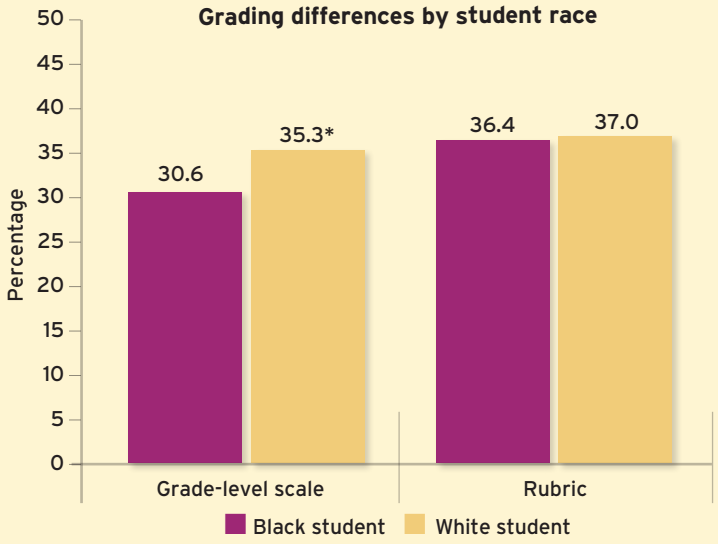
In looking at teachers by race, I find white teachers are approximately 8 percentage points less likely to rate the Black student's writing as being at grade-level or above compared to the white student's writing. By contrast, teachers of color do not show evidence of evaluation bias.

I also estimate grading bias among teachers who work in more and less racially diverse schools. As in my other analyses, bias is evident when teachers applied vague grade-level standards but not when they used specific criteria on a rubric. Teachers working in schools where no one particular race or ethnic group makes up a clear majority of enrolled students show the most bias in applying the grade-level scale. They are 13 percentage points less likely to rate the writing sample as on or above grade level if it was written by a Black student. There are no significant differences in the ratings assigned by teachers working in predominantly Black, Latinx, or white schools.

Some of these results raise important questions about student-teacher race and gender matching. Given that my teacher subgroup analyses are exploratory, and that this sample is not nationally representative (though it is national in scope), we cannot know whether these findings reflect patterns in the broader population. But they should inspire new hypotheses for further research. Female teachers showed racial bias in grading a Black male student compared to a white male student, whereas male teachers did not. Is a teacher less likely to exhibit racial bias against a student if the student shares the teacher's gender?

Rubrics Decrease Racial Bias in Grading Writing (Figure 2)

When teachers used a vague "grade-level" scale, they were 4.7 percentage points more likely to rate a white student's writing at grade-level or above compared to the same sample written by a Black student. However, when teachers used a rubric with specific criteria, the difference in grading for a white or Black student was no longer statistically significant.



NOTE: Figure shows the percentage of teachers rating the assignment "at grade level" or above (Grade-Level Scale) or "recounts an event with some detail" or above (Rubric). Estimates are adjusted for teacher gender, grade-level, race/ethnicity, experience, and school racial demographics.

* = difference in grades is statistically significant at the 95% confidence level.

SOURCE: Author's calculations

Limiting opportunities for biased decisions may have more immediate impact on equitable student evaluation than current forms of anti-bias training.

Results by Racial Attitudes

I also look at the relationships between teachers' grading bias and racial attitudes as measured by the implicit association test and explicit "warmth" questions.

Both tests showed attitudes that favor whites. The implicit attitudes test found that teachers had a significant association of white students as being more competent than Black students, by 41 percent of a standard deviation. The explicit "thermometer" questions measure showed a small and not significant preference for European Americans compared to Black Americans.

It stands to reason that teachers with higher measured levels of bias could show more bias on the grade-level evaluation measure. However, I find no relationship between teachers' measured attitudes and levels of grading bias, either on the vague grade-level scale or the specific rubric. In no case does the magnitude of the bias differ significantly by teachers' implicit or explicit racial attitudes.

There are several possible explanations for this. First, tests of implicit bias have limitations, and the implicit association test's validity as a test of individual attitudes has been questioned. Second, it may be that my sample size was too small to detect the true influence of implicit attitudes on grading bias.

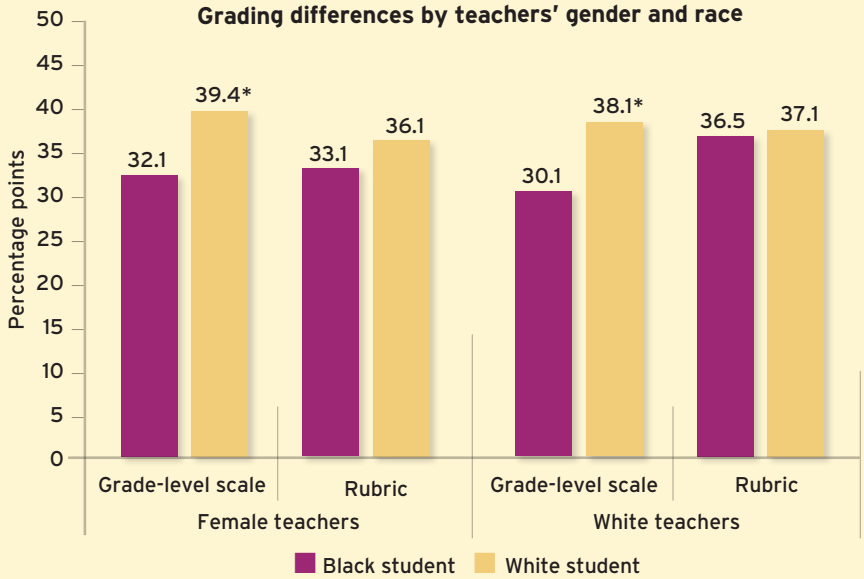
Indeed, this study did find some divergence in explicit and implicit attitudes. While the implicit attitudes test showed, on average, that teachers had a significant implicit association of white students as being more competent than Black students, the explicit measure showed a much smaller, and not significant, preference for whites. 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. Teachers may have been able to summon their explicit attitudes to dampen, but not entirely eliminate, the influence of implicit attitudes on their grading.

Implications

Scholars and decisionmakers have focused on two distinct approaches for mitigating the effects of negative

Larger Grading Differences for White and Female Teachers (Figure 3)

The differences in grades assigned to Black and white students when using the "grade-level" scale were largest for female teachers and white teachers. Even for these groups, there were no statistically different differences in grading when the teachers used the grading rubric with specific criteria.



NOTE: Figure shows the percentage of teachers rating the assignment "at grade level" or above (Grade-Level Scale) or "recounts an event with some detail" or above (Rubric). Estimates are adjusted for teacher gender, grade-level, race/ethnicity, experience, and school racial demographics.

* = difference in grades is statistically significant at the 95% confidence level.

SOURCE: Author's calculations

implicit racial attitudes: training programs that aim to reduce people's general implicit associations and efforts that engineer circumstances to reduce the impact that implicit stereotypes can have on a person's behaviors or judgments. My study shows strong potential from the latter approach when it comes to teacher grading. When teachers use a rubric that orients grading decisions to a limited number of specific, demonstrable criteria, they show no bias in their grading decisions. When teachers are asked to rate student work along a vaguer spectrum of performance, based on meeting "grade-level" standards, their grading favored the white student.

These findings raise a number of questions for future research. There may be something unique about the elements of this experiment that contribute to my results, whether the rubric, the writing sample, or the combination of the two. I can't rule out the possibility that the different sizes of each scale—seven points on the grade-level scale and four on the rubric—affected the amount of bias detected. We also should consider whether the impact of bias on grading could differ depending on the academic subject or the nature of the work being evaluated. The evaluation of student writing is likely more subjective than determining whether a student arrived at the correct answer to a math problem, for example. Additional experiments with other rating scales or other kinds of student work would be helpful.

The generalizability of these findings to the classroom is unknown. How might grading bias differ when teachers are grading their own students? Teachers' bias regarding students that they personally know may differ from the bias we find in this experimental setting, though past research has found some evidence of teachers' racial and gender-based bias being directed toward their own, familiar students. There may therefore be reason to recommend grading rubrics as a means to mitigate bias.

The present study does not offer direct evidence on whether rubrics would produce bias-reducing effects in school classrooms. It is possible that teachers hold strong student-specific biases that rubrics are less effective at overcoming. The findings in this study may be more generalizable to settings where raters are conducting anonymous reviews of essays in which the authors' identities may be signaled through context clues, such as state writing exams or SAT and GRE scoring.

I also note that, in this study, teachers were presented with a rubric without any training or examples on how it is appropriately applied. While using the rubric reduced bias, the grading task was simple and did not have any time constraints, unlike the complex evaluations of student work that teachers make as part of their day-to-day jobs. Previous research has suggested that rubrics do not improve grading reliability unless teachers are trained in how to use them. And in general, any efforts to reform or standardize teachers' classroom practices are less likely to succeed in the absence of aligned coaching and training.

I find white teachers are approximately 8 percentage points less likely to rate the Black student's writing as being at grade-level or above compared to the white student's writing. By contrast, teachers of color do not show evidence of evaluation bias.

Finally, policies that establish predetermined and clearly defined grading criteria may prove powerful in light of another finding from this study: that the overall bias was driven by white teachers and this bias seems to have been driven by an in-group preference among white teachers for white students. This finding aligns with calls to diversify the teaching force. Nationwide, 79 percent of teachers are white compared to 48 percent of students. 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, which suggests an ongoing disadvantage for students of color. Insofar as this imbalance powers biased evaluations of student work, it 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.

There are a great many ways in which racism, past and present, affects the educational opportunities of Black students. One proximate cause of inequality that school and district leaders have an opportunity to address is teachers' racially biased evaluations of students. A relatively simple tool may help start to mitigate the effects of teachers' racial biases on students.

David Quinn is assistant professor of education at the University of Southern California Rossier School of Education.