Received: 1 June 2021 | Accepted: 9 November 2021

DOI: 10.1002/aepp.13217

SUBMITTED ARTICLE

Mitigating implicit bias in student evaluations: A randomized intervention

Brandon Genetin[1] | **Joyce Chen**[1] | **Vladimir Kogan**[2] | **Alan Kalish**[3]

[1]Department of AED Economics, The Ohio State University, Columbus, Ohio, USA

[2]Department of Political Science, The Ohio State University, Columbus, Ohio, USA

[3]Office of Academic Affairs, The Ohio State University, Columbus, Ohio, USA

Correspondence
Brandon Genetin, Department of AED Economics, The Ohio State University, Columbus, OH, USA.
Email: genetin.4@osu.edu

Abstract

We conduct a randomized control trial to assess the efficacy of utilizing modified introductory language in student evaluations of instruction to mitigate implicit bias. Students are randomly assigned within courses to three treatment arms and shown so-called "cheap talk" scripts referencing implicit bias, the high stakes associated with student evaluations, and the combination of the two. We analyze both the impact assignment of the treatment has on completion rates as well as the effect on average instructor rating. Our analysis indicates assignment has statistically significant effects on the likelihood of response for those assigned the combined treatment, though the effects are heterogeneous with respect to both instructor and student race/ethnicity and gender. We further find the high-stakes treatment leads to higher average scores for racial/ethnic minority instructors with no significant effects from the implicit bias and combined scripts.

KEYWORDS

behavior, bias, education, gender, implicit bias

JEL CLASSIFICATION

J15, J16

Study Pre-Registration: AEA RCT Registry, AEARCTR-0006868

A growing body of research documents systematic differences in how students evaluate college instructors, with women, non-native English speakers, and racial/ethnic minorities receiving systematically lower ratings (Boring et al., 2016; Holman et al., 2019; Kreitzer & Sweet-Cushman, 2021). Given the weight placed on student evaluations in high-stakes reappointment, tenure, and promotion decisions, such biases in student evaluations could result in significant downstream disparities in the employment opportunities and career progression paths for members of these historically underrepresented groups.

Multiple studies report how women are unfavorably compared to men, often perceived as less adept, intelligent, or organized (Arbuckle & Williams, 2003; Boring et al., 2016; McPherson et al., 2009). This holds true even when course content is identical (Mengel et al., 2019) or the instructor's gender is experimentally manipulated (MacNell et al., 2015).

The scale of these biases vary. Research by Chavez and Mitchell find that when identical online courses taught by an undisclosed, similar instructor are perceived to be taught by different genders, female instructors receive evaluations that are 5.81% lower on average than male instructors (Chavez & Mitchell, 2020). Similarly, non-white instructors receive 3.94% lower average scores than their white colleagues. These biases are found to be predominantly driven by males, with classes comprised solely of men rating female instructors 0.207 SDs lower than male instructors, as compared to all-female classes rating female instructors 0.076 SDs lower (Mengel et al., 2019).

From a labor market perspective, these biases pose significant efficiency losses. While student evaluations are a noisy indicator of teaching quality, implicit bias in these evaluations on the basis of race and gender creates inequitable barriers for promotion. This, in turn, undermines diversity and representation in academia, which can then further confirm implicit biases. Moreover, women and men of color will have to devote more time to teaching in order to receive the same evaluation scores as their white male colleagues, which then detracts from their research productivity. Alternatively, instructors may be inclined to substitute more effective teaching approaches for those who earn higher evaluation scores, which may then compromise learning outcomes and student achievement.

While the literature on the presence of equity bias in student evaluations is plentiful, research into closing this gap is scant. Peterson and his coauthors conduct an experimental intervention designed to reduce gender bias in student evaluations of college instruction (Peterson et al., 2019). The intervention was carried out at Iowa State University and involved students taking introductory courses in American politics and biology. At the end of the semester, a randomly selected subset of the students in these courses completed the standard course evaluation survey (making up the control group), while the other half read a short prompt designed to mitigate gender biases prior to completing their evaluations (treatment group).

Peterson found that students assigned to the treatment group provided significantly higher ratings of female instructors compared to other students taught by the same instructors but who did not receive the prompt, with no impact on male faculty ratings. They further find that the improvement in the ratings of female instructors was driven exclusively by changes in the ratings of male students. We build on this study by implementing a randomized control trial at The Ohio State University. While previous research is limited to a few classes, our intervention was open to all instructors in the Colleges of Arts and Sciences and Food, Agricultural and Environmental Sciences, with roughly 400 instructors and over 800 classes electing to participate in the study. Additionally, our intervention includes three treatments to differentiate the impact of priming about implicit bias versus, or in addition to, language about the high-stakes nature of student evaluations.

In this paper, we focus on the study design and challenges in conducting randomization to maximize covariate balance across treatment groups. We augment this focus by taking a careful look at response rates among students (i.e., completion of the evaluation of instruction) as a function of treatment assignment. Our data suggest that the treatments do, in fact, have a statistically significant effect on the likelihood of response. However, the effects vary across treatment types and are heterogeneous with respect to both instructor and student race/ethnicity and gender.

After analyzing differences in completion rates, we additionally estimate the impact of treatments on average instructor ratings. We find the high stakes treatment leads to higher average scores for racial/ethnic minority instructors with no significant effects from the implicit bias and combined scripts. We attribute the lack of results to both changes in response rates induced by the scripts as well as the low variability in average scores. We conclude the paper with a brief discussion of policy implications for classroom utilization and possible reasons for the heterogeneous impact across treatments.

METHODS AND DATA

Intervention

We utilize "cheap talk" scripts that are shown to students when they enter the online system (supported by Bluenotes) for student evaluations of instruction (Cummings & Taylor, 1999). The language utilized by Peterson references both implicit bias related to race and gender as well as the high stakes of student evaluations for performance evaluation, promotion, and tenure. To disentangle these elements, we followed Peterson but utilized three separate treatment arms. In Treatment A, students are reminded only about implicit bias with the following language:

> The Ohio State University recognizes that student evaluations of teaching are often influenced by students' **unconscious** and **unintentional** biases about the race and gender of the instructor. Women and instructors of color are systematically rated lower in their teaching evaluations than white men, even when there are no actual differences in the instruction or in what students have learned.
> As you fill out the course evaluation, please keep this in mind and make an effort to resist stereotypes about professors. Focus on your opinions about the content of the course (the assignments, the textbook, the in-class material) and not unrelated matters (the instructor's appearance).

In Treatment B, students are reminded only about the high stakes associated with their evaluations, with the following language:

> Student evaluations of teaching play an important role in the review of faculty. Your participation in this process is essential; having feedback from as many students as possible provides a more comprehensive view of the strengths and weaknesses of each course offering, allowing instructors to improve their practices and increase learning. Moreover, your opinions influence the review of instructors that

takes place every year and will be taken into consideration for decisions regarding promotion and tenure.[1]

In Treatment C, students are shown both scripts together, and students in the control group are not shown any additional introductory language when accessing their evaluations of instruction.

Recruitment

The study was open to all faculty teaching undergraduate courses in the Spring 2021 term (second 7.5-week and full 15-week courses)[2] in the Colleges of Arts and Sciences and in the College of Food, Agricultural, and Environmental Sciences at The Ohio State University.[3] An invitation to participate was sent out via e-mail to all eligible instructors, and a reminder e-mail was sent 1 week prior to the closing of the 2-week enrollment period. E-mails were sent directly by the study team, utilizing instructor lists provided by the Office of Academic Affairs, as well as by the Office of Diversity and Inclusion and diversity officers in the participating colleges, in order to encourage the participation of instructors from historically underrepresented groups. Opting into the study opted in all courses taught by the instructor during the Spring 2021 semester. A waiver of consent for students was approved by the Institutional Review Board at Ohio State.

To minimize risk to instructors, expanded reports on student evaluations of instruction are provided to all participants. These reports provide evaluation scores disaggregated by treatment group, as well as average scores in comparable courses (based on size, level, and unit) disaggregated by treatment group. Additionally, department heads and deans are provided with guidance about how to interpret the impact of the intervention, as well as a discussion of the average impacts estimated by the study. In particular, the expanded reports seek to mitigate any potential backlash from the intervention, wherein students reminded of implicit bias may have the desire to retaliate against the instructor.

However, we also recognize that the intervention may have spillover effects on nonparticipating faculty, given that treatment occurs at the student rather than course level. There may also be leakage across peer groups if students discuss the content of the information treatment. Explanation of these issues to department heads and college deans will aid in the interpretation of student evaluations of instruction for the annual review and promotion and tenure processes.

Randomization

Students were randomly assigned to one of three treatment groups and a control group. Randomization was done at the student level and stratified by course, in order to mitigate concerns about the correlation between study participation and student evaluation scores. Two additional factors required consideration in conducting the randomization. First, within-course randomization of students across three treatment arms and one control group would create trivial size cells in smaller courses, particularly since the response rate for student evaluations is typically only 60%. Therefore, to make the results as informative as possible for individual instructors, courses with less than 40 students enrolled were randomized between the control group and

the implicit bias treatment only. In courses with 40 or more students, all three treatment scripts were given.

Second, the relatively high degree of overlap of students in multiple participating courses created concerns about contamination across groups. To mitigate these concerns, we manually reassigned students to receive the same treatment throughout if they were enrolled in multiple participating courses. To do this, the student was given the treatment that was randomly assigned in the largest participating course. For example, consider a student in three courses, Class 1 with 72 students, Class 2 with 43 students, and Class 3 with 18 students. The student was randomly assigned in each course as follows: Treatment B for Class 1, Treatment C for Class 2, and the Control for Class 3. Because Class 1 is this student's largest class, the student was assigned to Treatment B for all of their courses.[4] The randomization received in the largest course was maintained in order to optimize covariate balance, given that within-strata balancing is more difficult to achieve in smaller strata, and we did not want to compromise balance in the larger courses.

PARTICIPATION AND COVARIATE BALANCE

Instructor characteristics

Of the 2480 eligible instructors, 400 (16%) agreed to participate in the study.[5] The demographic breakdown of the instructors is shown in Table 1. Two-thirds of participating instructors are female, and women were significantly more likely to opt in than men. Almost three-fourths of participating instructors are white, and white instructors were significantly more likely to opt in, while Asian instructors were significantly less likely to opt in. We also observe significant differences in study participation by faculty type and rank. Assistant and Associate Professors and Senior Lecturers were more likely to opt in, while Graduate Teaching Associates and Lecturers were less likely to participate in the study. Similarly, instructors with greater years in rank/track were also less likely to opt in. The participating instructors teach 849 classes with a total enrollment of 33,975 students, comprised of 24,861 unique students. The size of these classes varies widely. The COVID-19 pandemic shifted many courses online, subsequently pushing introductory courses to combine sessions and contain as many as 1100 students. Conversely, there are roughly 9000 students in classes smaller than 40 people with the largest percentage of students (17%) in classes between the sizes of 20–29 people.

Student characteristics

Randomization produced 6100 students assigned to the control group, 8894 students assigned to the implicit bias treatment, 4889 students assigned to the high-stakes treatment, and the remaining 4886 students assigned to the combined treatment. The demographic composition of students in courses eligible for study inclusion is largely comparable to that for the universe of undergraduates at The Ohio State University, though eligible students are somewhat more likely to be female, non-white, and at a lower academic rank. This may be reflective of the prevalence of introductory general education courses in Arts and Sciences. Nonetheless, the sample of participating courses is largely comparable to the universe of undergraduate students, suggesting that results can be generalized across Colleges.[6]

TABLE 1 Comparison of instructor characteristics

	Participating	Not participating	*t*-statistic
Gender and age			
Female	66.3%	42.5%	8.87***
Age	41.3	40.7	0.75
Race and ethnicity			
American Indian	0.3%	0.2%	0.04
Asian	9.3%	13.2%	2.13**
Black	3.3%	3.8%	0.55
Hispanic	5.8%	5.9%	0.10
Two or more races	1.8%	2.3%	0.68
Undisclosed	6.0%	7.3%	0.94
White	73.6%	67.1%	2.54**
Job title			
Assistant professor	11.1%	4.8%	4.91***
Associate professor	17.3%	11.3%	3.38***
Professor	15.1%	12.1%	1.67*
Grad. teaching assoc.	20.6%	31.7%	4.46***
Lecturer	13.1%	17.3%	2.10**
Senior lecturer	8.8%	5.9%	2.20**
Time in job			
Years at university	8.1	8.3	0.27
Years in rank	5.9	9.1	5.01***
Years in track	11.9	16.1	4.79***
Observations	398	2082	-

Note: Reported *t*-Statistics are calculated using a two-sample *t*-test comparing group proportions and means. Significance of *t*-scores is reported at the 10% (*), 5% (**), and 1% (***) levels. Reported values for characteristics other than Age are the proportion of individuals who fall into that particular category. "Participating" includes all instructors who agreed to participate in the study. "Not Participating" includes all instructors who did not consent to the study but were eligible to participate.

Students in participating courses who actually completed the evaluation of instruction are also somewhat more likely to be female, though white students have slightly higher response rates. Students at lower academic ranks also have higher response rates, and the composition is similar to that for the population of students enrolled in the study. These differences suggest that response rates may have been affected by the intervention, and this may also differ across demographic groups.

Just over 60% of the students in our study were enrolled in only one participating course, and about one-third are enrolled in two or three participating courses.[7] Because student evaluations of instruction at our institution are hosted on a platform separate from the course management system, students tend to complete evaluations for all courses during the same session. Indeed, we can see that, among students in our study, those who complete an evaluation also complete evaluations for all participating courses. We also observe somewhat higher completion rates for students enrolled in more than three participating courses. Participating students see the same script preceding each evaluation, so enrollment in multiple participating courses is unlikely to increase the

student-level response rate. Rather, this difference is likely related to characteristics of students taking multiple courses in the two participating colleges (e.g., program of study and rank).

In Table 2, we see the average scores by subgroup for the response to Question 10 in the Student Evaluation of Instruction (SEI): "Overall, I would rate this instructor as ..." with "Poor" assigned a value of 1, "Fair" assigned a value of 2, "Neutral" assigned a value of 3, "Good" assigned a value of 4, and "Excellent" assigned a value of 5. In the control group, these scores are quite high on average at 4.38. Interestingly, white students seem to rate their instructors somewhat lower, as do students in lower academic ranks. Average SEI scores are significantly higher in the implicit bias treatment group, particularly for white male students, as well as

TABLE 2 Student evaluation of instruction average scores by student characteristics

	Implicit bias	High stakes	Combined	Control
Overall	4.42**	4.37	4.36	4.38
Student sex				
Female	4.43	4.42	4.39	4.41
Male	4.40**	4.31	4.32	4.35
Student race and ethnicity				
Asian	4.43	4.35	4.24***	4.42
Black	4.32	4.49	4.29	4.38
Hispanic	4.39	4.36	4.43	4.39
Two or more races	4.46	4.34	4.36	4.44
Undisclosed	4.39	4.43	4.43	4.35
White	4.42***	4.36	4.36	4.36
Student rank				
Freshman	4.28	4.25	4.11	4.21
Sophomore	4.33	4.27	4.30	4.33
Junior	4.44*	4.44	4.38	4.38
Senior	4.52	4.47	4.49	4.49
Course grades				
A	4.60**	4.54	4.56	4.55
A−	4.39	4.37	4.28	4.35
B	4.22	4.22	4.20	4.26
C	4.15	4.08	3.98	4.08
Non-Pass	4.02*	3.83	3.69	3.72
Pass	4.22**	4.09	4.15	3.98
Observations	7523	4734	4810	5492

Note: Significance of coefficients is reported at the 10% (*), 5% (**), and 1% (***) levels. Asterisks (*) indicate the mean of the said treatment group is significantly different from the mean of the control group. Reported SEI Scores are collected from the question: "Overall, I would rate this instructor as...*Poor, Fair, Neutral, Good, Excellent*." Values of "1" indicate an instructor rating of "Poor," a value of "2" for "Fair" and so on. "Implicit Bias" is the average SEI score for said student characteristic for students who received the implicit bias treatment script. "High Stakes," "Combined," and "Control" represent the same, albeit with their respective treatment scripts. Data for this analysis include all completed evaluations.

those at the extremes of the grade distribution. In the combined treatment group, Asian students gave significantly lower SEI scores, relative to the control group.

Balance tests

We present covariate balance tests for student characteristics across treatment and control groups.[8] Each cell reports the coefficient on the treatment variable (β) from the following regression specification:

$$\text{Characteristic}_s = \alpha_0 + \beta_1 \mathbf{Treatment_s} + \eta_c + \varepsilon_s,$$

where s represents each student, and η_c represents course-level fixed effects. Note also that this is a student-level regression, although some students participated in the study through multiple courses. "Treatment" represents an indicator for enrollment in a treatment group.

The student characteristics we test for balance include gender, race/ethnicity, and rank. Because our main focus is the average treatment effect across strata, we utilize this specification rather than an expanded specification allowing treatment coefficients to differ by strata (Firpo et al., 2020). For each treatment, only a handful of student characteristics exhibit a significant correlation with treatment status. Additionally, we conduct a test of joint significance by regressing treatment status on the full set of student characteristics. Based on this, we cannot reject the hypothesis that all characteristics jointly have zero effect on treatment status. Overall, covariates appear to be well-balanced for each treatment group, both individually and jointly, suggesting that randomization was successful.

SEI COMPLETION RATES

In this section, we examine the extent to which the intervention itself may have affected SEI response rates. This is critical to our understanding of how the intervention affects ratings as well. For example, students who have a negative perception of their course may be more likely to simply not complete the SEI after reading the treatment script, perhaps in recognition of the influence of implicit bias on their negative perception. Similarly, students in the high stakes and combined treatment groups may be more reluctant to provide negative feedback when reminded of these factors prior to viewing the SEI. While this may not be of primary concern to instructors, it does affect our understanding of how accurately SEI scores represent the true range of student experiences. And, to the extent that the intervention reduces completion, universities may need to consider potential trade-offs between maximizing response rates and mitigating implicit bias.

There are 24,861 students and 37,794 observations used in the analysis of SEI completion rates. Of the 24,861 students enrolled in participating courses, 14,353 students (57.7%) completed their SEIs. Accounting for students completing SEIs for multiple classes, there were 22,726 completed student evaluations out of the possible 37,794, a 60.1% response rate.

The use of a randomized control trial allows us to circumvent many identification issues around endogeneity. It is not clear a priori which direction this bias would run. Instructors who have received low evaluations in the past may be enthusiastic about participating in the study in hopes of mitigating implicit bias. Conversely, they may be reluctant to participate if they

anticipate that the treatment will, instead, induce resentment or animosity from students. By randomizing within courses, we can estimate the effect of the interventions, while holding fixed all of these unobserved instructor and course characteristics that may be correlated with both student response rates (evaluation scores) and the choice to participate in the study. Because we cannot observe if students who are assigned treatment scripts view those scripts, estimates for SEI completion analysis should be interpreted as intent to treat effects, that is, the effect of treatment *assignment* on the outcome.

We use the following linear probability model to assess the impact of treatment on SEI completion:

$$\text{Complete}_{s,c} = \alpha_0 + \alpha_1 \boldsymbol{X_s} + \alpha_2 \boldsymbol{X_{s,c}} + \alpha_3 \mathbf{Treat_s} + \beta(\mathbf{Treat_s} \times \mathbf{InstructorChar_c}) + \eta_c + \varepsilon_s,$$

where $\text{Complete}_{s,c}$ is a $\{0,1\}$ binary variable in which values of 1 indicate student s completed their SEI for course c and 0 if they did not.

All specifications include a course fixed effect η_c to account for unobserved characteristics of instructors and courses. Additional covariates include student characteristics ($\boldsymbol{X_s}$), consisting of a binary indicator for female students (Gender_s), and sets of indicator variables for student race (Race_s), rank (Rank_s), course of study (Major_s), and the student's final grade in the course ($\text{CourseGrade}_{s,c}$). Our primary variables of interest are a binary variable indicating whether the student is in a treatment (implicit bias, high stakes, and combined) group ($\mathbf{Treat_s}$), as well as interactions between the treatment variable and instructor's gender and race ($\mathbf{Treat_s} \times \mathbf{InstructorChar_c}$).

Of the 22,726 completed SEIs, 8,105 (35.7%) were completed using the mobile app. However, the mobile app did not have the capability to implement the treatment scripts into its interface. To keep our point estimates conservative, we assign those completing SEIs via the mobile app to their original treatment groups, in case students viewed the treatment scripts via the web platform before responding to the survey on the mobile app. This seems unlikely because student evaluations are accessed through a separate linked portal and are generally completed in one session. As an additional robustness check, responses via the mobile app are dropped from the analysis in an appendix table. The point estimates remain largely unchanged in sign and magnitude.[9]

Results from our regression analysis are shown in Tables 3 and 4, examining response rates for each treatment arm individually. For ease of interpretation, net effects are also reported as linear combinations of the direct effect of treatment with the relevant interaction effect(s) and the associated SE. Given that over two-thirds of instructors and students identify as white, we do not have enough data or statistical power to estimate effects for each race/ethnicity group individually. Instead, students and instructors who identify as American Indian, Asian, African American, Hispanic, or Native Hawaiian are grouped into the "racial and ethnic minority" (REM) variable. Instructors who identify as "Two or More Races" or "Undisclosed" (roughly 7.8% of instructors) were individually assigned to "REM" or "White" based on manual review of faculty profiles confirmed by multiple evaluators.[10]

In Table 3, we see that, overall, receiving the treatment scripts has no statistically significant effect on the likelihood of completing the SEI. With regard to student characteristics, we see that women are between 8 and 9 percentage points (ppt) more likely to complete their SEIs relative to men. These results are consistent throughout treatment arms and specifications. We further find consistent results for underclassmen who are 3–5 ppt more likely to complete evaluations than upperclassmen. Interestingly, students who receive a low or failing grade are 32–33 ppt less likely to complete student evaluations than students who receive a high grade,

TABLE 3 Linear probability model: an indicator of student evaluation of instruction completion (*no interaction terms*)

	Implicit bias	High stakes	Combined
Treatment			
Treatment script	0.000 (0.007)	0.010 (0.007)	−0.001 (0.007)
Student sex			
Female	0.084*** (0.007)	0.090*** (0.008)	0.080*** (0.008)
Student rank			
Underclassmen	0.034*** (0.009)	0.041*** (0.010)	0.048*** (0.010)
Student race			
Asian	−0.027** (0.013)	−0.037** (0.014)	−0.016 (0.014)
African American	−0.045*** (0.013)	−0.048*** (0.015)	−0.044*** (0.015)
Hispanic	−0.018 (0.015)	−0.024 (0.016)	−0.022 (0.016)
Nonresident Alien	0.023 (0.016)	0.024 (0.017)	0.056*** (0.017)
Undisclosed	0.008 (0.020)	0.004 (0.023)	0.004 (0.024)
Two or more races	−0.029* (0.016)	−0.031* (0.018)	−0.027 (0.018)
Course grade			
Low/failing grade	−0.316*** (0.012)	−0.324*** (0.014)	−0.326*** (0.014)
Included covariates			
Course FE	✓	✓	✓
Program of study	✓	✓	✓
$\overline{R}^2$	0.140	0.144	0.135
Observations	20,924	16,230	16,426
Constant	0.567***	0.543***	0.533***

Note: Significance of estimates is reported at the 10% (*), 5% (**), and 1% (***) levels. "Treatment" values should be interpreted relative to the control group. Parentheses indicate robust SEs for coefficient estimates. The analysis utilizes the original treatments of mobile completions. Student ranks are interpreted relative to "Upperclassmen." "Underclassmen" includes freshmen and sophomores while "Upperclassmen" includes juniors and seniors. Student races are interpreted relative to "White." Course grades are interpreted relative to grade "High Grades." "High Grades" include students who received an A(−), B(+/−), or C(+/−). "Low/Failing Grades" include students who failed the course, received a "Pass" in a Pass/Fail course, an Emergency Pass, or an Incomplete.

going against conventional wisdom that students who perform poorly in a class are more likely to complete their evaluations. Last, we find that Asian (3–4 ppt) and African American (4–5 ppt) students are less likely to complete evaluations, as are students identifying with two or more races (3 ppt).

In Table 4, we see that treatments have heterogeneous effects related to both instructor and student gender as well as race/ethnicity, and these also differ across treatment groups. Under the implicit bias treatment, female students are significantly less likely to complete course evaluations, though only for male instructors. Interestingly, we do not observe the same effect for the combined treatment, even though it also contains language about implicit bias. Under the high-stakes treatment, students in racial/ethnic minority groups are significantly less likely to complete course evaluations but, again, this applies only to male instructors. We also do not see

TABLE 4 Linear probability model: an indicator of student evaluation of instruction completion (*all interaction terms*)

	Implicit bias	High stakes	Combined
Treatment			
Treatment script	0.025 (0.022)	0.003 (0.024)	−0.008 (0.025)
Instructor effects			
Treatment × Female	−0.031 (0.028)	0.007 (0.031)	0.023 (0.031)
Treatment × REM	0.014 (0.046)	0.036 (0.052)	−0.106** (0.053)
Treatment × Female × REM	0.030 (0.058)	0.011 (0.065)	0.120* (0.067)
Student effects			
Treatment × Student Female	−0.055* (0.030)	0.037 (0.032)	0.006 (0.033)
Treatment × Student REM	−0.045 (0.047)	−0.100* (0.051)	0.029 (0.051)
Treatment × Student Female × Student REM	0.042 (0.043)	0.043 (0.050)	−0.091* (0.048)
Female instructor interactions			
× Treatment × Student Female	0.061** (0.025)	−0.025 (0.031)	−0.043 (0.031)
× Treatment × Student REM	0.012 (0.039)	0.088* (0.049)	−0.136*** (0.048)
× Treatment × Student Female × Student REM	−0.076 (0.053)	−0.101 (0.064)	0.173*** (0.063)
REM instructor interactions			
× Treatment × Student Female	0.003 (0.043)	−0.039 (0.058)	0.121** (0.059)
× Treatment × Student REM	−0.089 (0.061)	0.101 (0.077)	0.023 (0.082)
× Treatment × Student Female × Student REM	−0.002 (0.087)	−0.091 (0.120)	−0.124 (0.120)
Female and REM instructor interactions			
× Treatment × Student Female	0.034 (0.054)	−0.018 (0.071)	−0.106 (0.072)
× Treatment × Student REM	0.107 (0.080)	−0.127 (0.096)	0.148 (0.101)
× Treatment × Student Female × Student REM	0.009 (0.110)	0.191 (0.145)	−0.068 (0.144)
Student rank and interactions			
Underclassmen	0.028** (0.012)	0.039*** (0.012)	0.039*** (0.012)
Treatment × Underclassmen	0.008 (0.014)	0.006 (0.015)	0.019 (0.015)
Course grade and interactions			
Low/failing grade	−0.321*** (0.018)	−0.315*** (0.018)	−0.313*** (0.018)
Treatment × low/failing grade	0.009 (0.023)	−0.022 (0.027)	−0.031 (0.027)
Student sex			
Female	0.091*** (0.023)	0.085*** (0.023)	0.089*** (0.024)
Student race			
Asian	0.013 (0.037)	0.002 (0.037)	0.023 (0.037)
African American	−0.005 (0.036)	−0.005 (0.037)	−0.002 (0.037)
Hispanic	0.024 (0.037)	0.017 (0.037)	0.021 (0.038)

(Continues)

TABLE 4 (Continued)

	Implicit bias	High stakes	Combined
Nonresident Alien	0.042 (0.044)	0.046 (0.044)	0.078* (0.044)
Undisclosed	0.029 (0.045)	0.029 (0.046)	0.026 (0.047)
Two or more races	−0.003 (0.044)	−0.004 (0.045)	−0.002 (0.045)
Course FE	✓	✓	✓
Program of study	✓	✓	✓
$\overline{R}^2$	0.141	0.144	0.137
Observations	20,828	16,180	16,382
Constant	0.557***	0.535***	0.5233***

Note: Significance of estimates is reported at the 10% (*), 5% (**), and 1% (***) levels. "REM" represents Racial and Ethnic Minorities and includes American Indians, Asians, African Americans, Hispanics, and Native Hawaiians. "REM" values should be interpreted relative to the "White" category. "Treatment" values should be interpreted relative to the control group. Parentheses indicate robust SEs for coefficient estimates. Columns utilize the original treatments of mobile completions. Student ranks are interpreted relative to "Upperclassmen." "Underclassmen" includes freshmen and sophomores while "Upperclassmen" includes juniors and seniors. Course grades are interpreted relative to grade "High Grade." "High Grades" include students who received an A(−), B(+/−), or C(+/−). "Low/Failing Grades" include students who failed the course, received a "Pass" in a Pass/Fail course, an Emergency Pass, or an Incomplete. Student races are interpreted relative to "White."

a comparable effect with the combined treatment, despite the fact that it also includes language about the stakes associated with student evaluations. However, for the combined treatment, we find that students are less likely to complete evaluations for racial/ethnic minority instructors, though this applies only to male students. Moreover, male students of color are less likely to complete evaluations for female instructors. Treatment does not differentially impact response rates by student rank or course grade.

Taken together, our findings indicate that prompts about implicit bias, when combined with the high-stakes treatment, tend to reduce response rates, particularly among students of color, and specifically for male racial/ethnic minority instructors. Under the combined treatment, female instructors receive even fewer responses from male students of color, while minority instructors have higher response rates from female students. The implicit bias treatment alone reduces response rates from female students for male instructors, while the high-stakes treatment alone reduces response rates from minority students, again particularly for male instructors. This suggests that the treatments tended to discourage racial/ethnic minority students from completing evaluations of instruction for minority instructors, with the opposite effect on female students. Note that, given the inclusion of course fixed effects, we cannot discern whether these findings imply a net reduction in responses from minority students or only a reduction relative to white students. Nonetheless, this shift in the distribution of response rates across demographic groups, for both instructors and students, is an important factor for interpreting any effects, or lack thereof, on average evaluation scores.

SEI AVERAGE SCORES

Next, we examine responses to the summary evaluation question, "Overall, I would rate this instructor as ..." We focus on this question because it is the primary metric highlighted in reports generated for instructors and utilized for performance evaluation, including promotion

TABLE 5 Student evaluation of instruction average score analysis (*no interactions*)

	Implicit Bias	High Stakes	Combined
Treatment			
Treatment script	−0.017 (0.020)	−0.020 (0.023)	−0.021 (0.023)
Mobile indicator			
Mobile	−0.079*** (0.025)	−0.092*** (0.026)	−0.091*** (0.026)
Mobile × Treatment	0.021 (0.032)	0.058 (0.037)	0.064* (0.037)
Mobile indicator			
Female	0.015 (0.017)	0.031 (0.020)	0.013 (0.020)
Student rank			
Underclassmen	−0.021 (0.021)	−0.019 (0.024)	−0.035 (0.024)
Student race			
Asian	0.056* (0.031)	0.053 (0.035)	0.014 (0.034)
African American	−0.034 (0.034)	0.058 (0.039)	0.032 (0.039)
Hispanic	0.001 (0.036)	−0.031 (0.041)	0.044 (0.041)
Nonresident Alien	0.173*** (0.036)	0.163*** (0.040)	0.155*** (0.040)
Undisclosed	0.043 (0.047)	0.038 (0.054)	0.066 (0.056)
Two or more races	0.040 (0.039)	0.003 (0.045)	0.005 (0.045)
Course grade			
Low/failing grade	−0.263*** (0.039)	−0.329*** (0.046)	−0.307*** (0.047)
Included covariates			
Course FE	✓	✓	✓
Program of study	✓	✓	✓
$\overline{R}^2$	0.250	0.244	0.244
Observations	12,644	10,027	10,092
Constant	4.521***	4.447***	4.370***

Note: Significance of estimates is reported at the 10% (*), 5% (**), and 1% (***) levels. Parentheses indicate robust SEs for coefficient estimates. "Treatment" values should be interpreted as the impact of the specified treatment relative to the control group. Columns (1)–(3) represent results for the Implicit Bias, High Stakes, and Combined treatments, respectively. Student ranks are interpreted relative to "Upperclassmen." "Underclassmen" includes freshmen and sophomores while "Upperclassmen" includes juniors and seniors. Student races are interpreted relative to "White." Course grades are interpreted relative to grade "High Grades." "High Grades" include students who received an A(−), B(+/−), or C(+/−). "Low/Failing Grades" include students who failed the course, received a "Pass" in a Pass/Fail course, an Emergency Pass, or an Incomplete.

and tenure decisions, at The Ohio State University. We use the following ordinary least squares regression to assess the impact of treatment on instructor scores:

$$\text{Score}_{s,c} = \alpha_0 + \alpha_1 \boldsymbol{X_s} + \alpha_2 \boldsymbol{X_{s,c}} + \alpha_3 \mathbf{Treat_s} + \beta(\mathbf{Treat_s} \times \mathbf{InstructorChar_c}) + \eta_c + \varepsilon_s,$$

where $\text{Score}_{s,c}$ is a $1-5$ ordinal variable as described in Table 4 for indicating the score student s rated the instructor for course c. Covariates mimic those included in the linear probability models.

TABLE 6 Student evaluation of instruction average score analysis (*student gender and race interactions*)

	Implicit bias	High stakes	Combined
Treatment			
Treatment script	0.027 (0.055)	−0.016 (0.062)	0.001 (0.063)
Mobile indicator			
Mobile	−0.079*** (0.025)	−0.091*** (0.026)	−0.092*** (0.026)
Mobile × Treatment	0.020 (0.033)	0.058 (0.037)	0.065* (0.037)
Instructor effects			
Treatment × Female	−0.001 (0.068)	0.021 (0.077)	0.004 (0.078)
Treatment × REM	−0.134 (0.11)	−0.166 (0.126)	−0.047 (0.140)
Treatment × Female × REM	0.164 (0.139)	0.268* (0.157)	−0.019 (0.169)
Student effects			
Treatment × Student Female	−0.066 (0.072)	−0.015 (0.080)	−0.034 (0.081)
Treatment × Student REM	0.047 (0.118)	0.051 (0.137)	0.091 (0.129)
Treatment × Student Female × Student REM	0.082 (0.106)	0.134 (0.132)	−0.092 (0.118)
Female instructor interactions			
× Treatment × Student Female	0.049 (0.060)	0.047 (0.074)	0.066 (0.074)
× Treatment × Student REM	−0.004 (0.101)	0.054 (0.129)	0.101 (0.120)
× Treatment × Student Female × Student REM	−0.103 (0.132)	−0.248 (0.164)	−0.127 (0.155)
REM instructor interactions			
× Treatment × Student Female	−0.090 (0.104)	−0.168 (0.135)	0.018 (0.145)
× Treatment × Student REM	−0.119 (0.166)	−0.205 (0.192)	−0.180 (0.208)
× Treatment × Student Female × Student REM	0.105 (0.230)	0.734** (0.300)	−0.101 (0.304)
Female & REM instructor interactions			
× Treatment × Student Female	0.018 (0.131)	−0.048 (0.167)	−0.041 (0.175)
× Treatment × Student REM	0.034 (0.207)	0.013 (0.237)	0.371 (0.250)
× Treatment × Student Female × Student REM	0.048 (0.281)	−0.222 (0.358)	0.031 (0.360)
Included covariates			
Student demographics	✓	✓	✓
Course grades	✓	✓	✓
Course FE	✓	✓	✓
Program of study	✓	✓	✓
$\overline{R}^2$	0.250	0.245	0.244
Observations	12,580	9990	10,062
Constant	4.486***	4.430***	4.343***

Note: Significance of estimates is reported at the 10% (*), 5% (**), and 1% (***) levels. "Treatment Script" values should be interpreted as the impact of the specified treatment relative to the control group. Columns (1)–(3) represent results for the Implicit Bias, High Stakes, and Combined treatments, respectively. "REM" represents Racial and Ethnic Minorities and includes American Indians, Asians, African Americans, Hispanics, and Native Hawaiians. Parentheses indicate robust SEs for coefficient estimates. Student demographics include student sex, student race, and student ranks.

TABLE 7 Student evaluation of instruction average score analysis (*student rank interactions*)

	Implicit bias	High stakes	Combined
Treatment			
Treatment script	−0.024 (0.024)	−0.041 (0.028)	−0.032 (0.028)
Mobile indicator			
Mobile	−0.080*** (0.025)	−0.094*** (0.026)	−0.093*** (0.026)
Mobile × Treatment	0.022 (0.033)	0.063* (0.037)	0.066* (0.037)
Treatment interactions			
×Underclassmen	0.015 (0.032)	0.046 (0.037)	0.025 (0.037)
Included covariates			
Student demographics	✓	✓	✓
Course grades	✓	✓	✓
Course FE	✓	✓	✓
Program of study	✓	✓	✓
$\overline{R}^2$	0.250	0.244	0.244
Observations	12,644	10,027	10,092
Constant	4.524***	4.457***	4.375***

Note: Significance of estimates is reported at the 10% (*), 5% (**), and 1% (***) levels. "Treatment" values should be interpreted as the impact of the specified treatment relative to the control group. Parentheses indicate robust SEs for coefficient estimates. Columns (1)–(3) represent results for the Implicit Bias, High Stakes, and Combined treatments, respectively. "REM" represents Racial and Ethnic Minorities and includes American Indians, Asians, African Americans, Hispanics, and Native Hawaiians. Student demographics include student sex, student race, and student ranks. Student ranks are interpreted relative to "Upperclassmen." "Underclassmen" includes freshmen and sophomores while "Upperclassmen" includes juniors and seniors.

In Table 5, we see that the treatment scripts overall have no statistically significant impact on average evaluation scores, just as we found no impact on completion rates. Interestingly, we see that female students and underclassmen do not give significantly higher instructor ratings, on average, though they are more likely to complete the evaluations. Conversely, students receiving low or failing grades, who have lower response rates, give significantly lower scores.

In Tables 6–8, we again incorporate interactions with student and instructor characteristics to allow heterogeneous effects across race/ethnicity and gender groups. However, unlike with response rates, we find almost no significant effects of treatment on average ratings. The notable exceptions are that, under the high-stakes treatment, female minority instructors receive significantly higher scores, and all minority instructors receive significantly higher scores from female minority students. Treatment scripts, however, do not have any significant differential effects on instructor ratings for underclassmen and students with low or failing grades, consistent with our findings on response rates.

Altogether, our findings suggest that reminding students about the high stakes associated with their course evaluations leads to higher average scores for racial/ethnic minority instructors, particularly women. This appears to be driven in part by higher ratings from female minority students, suggesting some kind of affinity effect. However, we must also acknowledge the possibility that the lack of significant effects on average scores for the implicit bias and combined treatments may be due to changes in response rates induced by those treatments. Again,

TABLE 8 Student evaluation of instruction average score analysis (*course grade interactions*)

	Implicit bias	High stakes	Combined
Treatment			
Treatment script	−0.021 (0.020)	−0.021 (0.023)	−0.024 (0.023)
Mobile indicator			
Mobile	−0.078*** (0.025)	−0.091*** (0.026)	−0.090*** (0.026)
Mobile × treatment	0.020 (0.033)	0.058 (0.037)	0.062* (0.037)
Treatment interactions			
×Low/failing grade	0.102 (0.076)	0.032 (0.088)	0.091 (0.089)
Included covariates			
Student demographics	✓	✓	✓
Course grades	✓	✓	✓
Course FE	✓	✓	✓
Program of study	✓	✓	✓
$\overline{R}^2$	0.250	0.244	0.244
Observations	12,644	10,027	10,092
Constant	4.523***	4.447***	4.371***

Note: Significance of estimates is reported at the 10% (*), 5% (**), and 1% (***) levels. "Treatment" values should be interpreted as the impact of the specified treatment relative to the control group. Parentheses indicate robust SEs for coefficient estimates. Columns (1)–(3) represent results for the Implicit Bias, High Stakes, and Combined treatments, respectively. "REM" represents Racial and Ethnic Minorities and includes American Indians, Asians, African Americans, Hispanics, and Native Hawaiians. Student demographics include student sex, student race, and student ranks. Course grades are interpreted relative to grade "High Grades." "High Grades" include students who received an A(−), B(+/−), or C(+/−). "Low/Failing Grades" include students who failed the course, received a "Pass" in a Pass/Fail course, an Emergency Pass, or an Incomplete.

due to the inclusion of course fixed effects, it is unclear whether treatments induced higher response rates from white students or lower response rates from minority students, but this relative shift may have offset the effects of the intervention on scores. For example, if the true effect of the intervention on scores is positive, but white students respond more and give lower scores compared to minority students, it would appear that the intervention had no effect on net.

CONCLUSION

In this paper, we present details regarding the implementation of a randomized control trial at The Ohio State University to test the efficacy of "cheap talk" scripts in improving the information obtained from student evaluations of instruction. We explore two alternative prompts addressing (1) implicit bias related to race/ethnicity and gender and (2) the high stakes associated with student evaluations in the promotion and tenure decisions, as well as the combination of the two. Issues to consider in this type of randomization include contamination across treatment groups due to the enrollment of students in multiple participating courses, as well as the link between class size and statistical power within courses, given the need to minimize risks of research participation for instructors. The ways in which we addressed these issues do

not appear to have had significant effects on covariate balance across treatment groups, suggesting that they are feasible remedies.

Cognizant of the trade-offs between minimizing implicit bias versus discouraging participation in the process of evaluating instruction, we analyzed the effect of treatment scripts on both completion rates and average instructor scores. We found the treatments scripts tended to discourage racial/ethnic minority students from completing evaluations of instruction for minority instructors, with the opposite effect on female students. Meanwhile, results from the average score analysis suggested that reminding students about the high stakes associated with their course evaluations leads to higher average scores for racial/ethnic minority instructors, particularly women. We do not observe the same effect on scores with the implicit bias and combined scripts, suggesting that changes in response rates induced by those scripts may have been offsetting. Last, given the inclusion of course fixed effects, we can not discern whether significant results imply net reductions/increases or reductions/increases relative to white students.

The lack of significant effects on average scores for the implicit bias and combined treatments may also be attributed to the low variability in SEI scores. Since instructors are only rated on a 5-point scale and traditionally have average ratings of 4 or higher, the true effects of treatment scripts may be minimized in the average score analysis. Instead, treatments may affect the likelihood of receiving a low (1 or 2) or high (5) score from a student. More research on the tails of the distribution is needed to confirm this hypothesis.

All three treatments are found to have statistically significant negative effects on SEI completion rates for male racial/ethnic minority students, with the opposite effect for female students, both white and minority. Female instructors also tended to have higher response rates overall when students were presented with any of the three scripts. It is notable, then, that minority and female instructors receive higher scores only with the high-stakes intervention. If we assume that the implicit bias script has a weakly positive effect on course evaluation scores, then the relatively lower response rates due to the implicit bias and combined scripts suggest that prompts about implicit bias may reduce (increase) the prevalence of high (low) scores from female and minority (male and white) students, leading to no change in average scores. Alternatively, if the implicit bias script has a negative effect on scores, then the net-zero impact on scores could be driven by the reduced (increased) prevalence of low (high) scores from female and minority (male and white) students.

With regard to the continued inclusion of such "cheap talk" scripts in student evaluations of instruction, it is clear that these interventions have no negative effect on instructor average ratings. This helps to alleviate concerns about backlash, so such scripts could be implemented widely without fear of causing harm to instructors' course evaluation ratings. Moreover, female and racial/ethnic minority instructors, who have been shown to receive lower scores due to bias, experience a significant increase in average scores.

Prior to scaling up this intervention, however, more research on the impact of treatments on response rates is warranted. It will be important to discern whether response rates are higher overall, in addition to our analysis exploring relative response rates across student demographic groups. Additionally, qualitative research is needed to assess student responses to these scripts and whether the prompts are leading to reduced expression of implicit bias or curtailing the expression of constructive concerns with courses and instructors. While it may be appropriate to discourage students from expressing biased opinions, universities should exercise caution to ensure that students do not feel deterred from expressing genuine concerns. Given the importance of student evaluations of instruction in the promotion and tenure decisions and, ideally,

student learning and achievement, the issues presented in this paper must continue to be investigated carefully and systematically.

ACKNOWLEDGMENTS

We thank BlueNotes and the Office of Diversity and Inclusion at Ohio State for their generous support. We also like to thank Nick Magnan (The University of Georgia) for his helpful suggestions.

ENDNOTES

[1] Note, however, that students may not have a good understanding of the significance of promotion and tenure in this context, so this intervention may not provide full information about the stakes associated with their evaluations.

[2] We collect data after the initial COVID shock and randomize within courses to account for unobserved differences across instructors and courses that may affect student evaluations directly. However, it is possible that COVID continued to disproportionately impact certain types of instructors. Our analysis would understate (overstate) the impact of the intervention if, for example, implicit bias is exacerbated (reduced) by virtual instruction.

[3] The study was limited to these two colleges due to the investigators' familiarity with the role of student evaluations of instruction in the promotion and tenure processes in these units.

[4] For students in multiple classes with at least one above the size of 40, they maintain the treatment assigned to them in the largest class, even if the course has fewer than 40 students.

[5] Because we only had demographic data on 398 of the instructors, we limit our analysis to those 398 individuals

[6] Comparison of student characteristics is reported in Table S1.

[7] Table S2 displays the number of students enrolled in multiple courses participating in our studies, as well as the frequency with which these students completed evaluations.

[8] Results of the covariate balance test are reported in Table S3.

[9] See Table S4.

[10] Alternative specifications in which students who are categorized as "undisclosed" are moved to "REM" are reported in Table S5.

REFERENCES

Arbuckle, Julianne, and Benne D. Williams. 2003. "Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations." *Sex Roles* 49(9): 507–16.

Boring, Anne, Kellie Ottoboni, and Philip B. Stark. 2016. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." *ScienceOpen Research* .

Chávez, Kerry, and Kristina M. Mitchell. 2020. "Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity." *PS: Political Science & Politics* 53(2): 270–4.

Cummings, Ronald G., and Laura O. Taylor. 1999. "Unbiased Value Estimates for Environmental Goods: a Cheap Talk Design for the Contingent Valuation Method." *American Economic Review* 89: 649–65.

Firpo, Sergio, Miguel Foguel, and Hugo Jales. 2020. "Balancing Tests in Stratified Randomized Controlled Trials: A Cautionary Note." *Economics Letters* 186: 108771.

Holman, Mirya, Ellen Key, and Rebecca Kreitzer. 2019. "Evidence of Bias in Standard Evaluations of Teaching". http://www.rebeccakreitzer.com/bias/

Kreitzer, Rebecca, and Jennie Sweet-Cushman. 2021. "Evaluating Student Evaluations of Teaching: A Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform." *Journal of Academic Ethics*: 1–12.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2015. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40(4): 291–303.

McPherson, Michael A., R. Todd Jewell, and Myungsup Kim. 2009. "What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes." *Eastern Economic Journal* 35(1): 37–51.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. 2019. Gender bias in teaching evaluations. *Journal of the European Economic Association* 17(2): 535–566.

Peterson, David A., Lori A. Biederman, David Andersen, Tessa M. Ditonto, and Kevin Roe. 2019. "Mitigating Gender Bias in Student Evaluations of Teaching." *PLoS One* 14(5): e0216241.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Genetin, Brandon, Joyce Chen, Vladimir Kogan, and Alan Kalish. 2021. "Mitigating implicit bias in student evaluations: A randomized intervention." *Applied Economic Perspectives and Policy* 1–19. https://doi.org/10.1002/aepp.13217