Implicit Bias in Student Evaluations: A Pilot Study

Brandon Genetin, Joyce Chen, Vladimir Kogan, Alan Kalish

The Ohio State University

Abstract

A growing body of research documents systematic differences in how students evaluate college instructors, with women, non-native English speakers, and minorities receiving systematically lower ratings. 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. Motivated in large part by this concern, we seek to assess the efficacy of utilizing modified introductory language to mitigate implicit bias in student evaluations of instruction. Utilizing an RCT framework, so-called "cheap talk" scripts are randomly assigned within classes to survey respondents and describe the hypothetical biases that tend to arise in the study context. This is potentially a highly cost-effective strategy to improve the quality of information generated by student evaluations of instruction, while also minimizing inequities for underrepresented populations.

Keywords: *behavior, education, gender, bias, implicit bias*
JEL Codes: *J15, J16*
Study Pre-Registration: AEA RCT Registry, *AEARCTR-0006868*

Introduction

Research into student evaluations of college instructors has documented large systematic differences between the ratings of male and female instructors (Boring, Ottoboni, and Stark, 2016). Analysis from the last decade reveals these disparities are often driven by the evaluations of male students, though differences still exist in female student evaluations (Mengel et al, 2019). These differences hold true even when course and learning experiences are otherwise identical —including when instructor gender and race (as perceived by the students) are experimentally manipulated (Chavez and Mitchell, 2020). Similar results are found by Macnell, Driscoll and Hunt (2015) who utilized an online class format to let assistant instructors operate under two different gender identities. Results indicated that students rated male identities significantly higher than the female identity, regardless of the instructor's actual gender.

Motivated in large part by this concern, Peterson et al. (2019) report the results of an experimental intervention designed to reduce gender biases in student evaluations of college instruction. 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 et al. (2019) 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 the ratings of male faculty. They further find that the improvement in the ratings of female instructors were driven exclusively by changes in the ratings of male students.

This study seeks to build on the research of Peterson et al. (2019) by understanding the disparate outcomes in student evaluations that exists between white male instructors and females and minorities. Specifically, we look to measure the impact priming students with "cheap talk" scripts has on mitigating the systematic differences in student evaluations of instructors.

Research Design

HYPOTHESES: Our main hypothesis is that the treatment scripts will decrease the main effect of race and/or gender on student evaluation scores. This primarily should occur through the increase in scores for women and persons of color. We believe the treatment scripts may not increase the student evaluation scores for other instructors if our belief of a shrinking race and/or gender effect holds. To measure this, our main outcomes of interests are the average scores on the SEI questions. Since the purpose of the study is to improve the existing SEI instrument, we will use the same 10 questions currently asked:

1. The subject matter of this course was well organized
2. The instructor is well prepared
3. The instructor communicated the subject manner clearly
4. The instructor was genuinely interested in teaching
5. The instructor was genuinely interested in helping students
6. The instructor created an atmosphere conducive to learning
7. The course was intellectually stimulating
8. The instructor encouraged students to think for themselves
9. I learned a great deal from the instructor
10. Overall, I would rate this instructor as… *Poor, Fair, Neutral, Good, Excellent*

Responses on questions 1 through 9 range from “Strongly Disagree" to “Strongly Agree."

IDENTIFICATION STRATEGY: We will utilize a randomized control treatment to identify the differences in treatment and control groups. Randomization will occur within class sections in order to directly compare the groups to one another.

INTERVENTION: The study was open to all faculty teaching undergraduate courses in the Spring 2021 term (second 7.5-week and full 15-week courses) in the Colleges of Arts and Sciences and in the College of Food, Agricultural, and Environmental Sciences. An invitation was sent out via e-mail to all instructors who fit the requirements stated previously and was open for two weeks. Opting into the study also opted in all classes and students said instructor taught that semester.

Students were randomized into one of four groups. If a student was assigned to a treatment group, they saw one of the following texts at the top of the SEI, before answering the questions:

1. **High Stakes Treatment Script:** “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.”

2. **Implicit Bias Treatment Script:** "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)."

3. **Combined Treatment Script:** "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.

   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)."

4. **Control:** No treatment script.

Randomization occurred by assigning each student a random number in STATA and grouping numbers into each treatment or control group. To maximize efficiency and the usefulness of the results for instructors, the number of treatments assigned was a function of class size. If a class had less than forty students, only the implicit bias treatment was given. If a class had forty or more students, all three treatment scripts were given. To keep the coefficient estimates unbiased, we made sure students received the same treatment throughout if they were in multiple classes. To do this, the student was given the same treatment that was randomly assigned to them in their largest participating class. For example, if a student was in two classes, one with 72 students and one with 23 students, the student was randomly assigned a treatment in the class with 72 students and given the same treatment for the class with 23 students.

However, we recognize that in order to make sure students receive the same treatment throughout, there will be a sizeable possibility that students in classes smaller than 40 will have a treatment that is not the implicit bias one. This is because there is no clean way to both ensure there is only one treatment with classes under a size of 40 and three treatments for classes over the size of 40 as well as ensure students in multiple classes have a consistent treatment or control. So, in order to maximize instructor usefulness, if a student is only in classes under the size of 40 they will either have the implicit bias treatment or the control treatment. If a student is in multiple classes with at least one above the size of 40, they will have the treatment assigned to them in the largest class.

Data

The two-week opt-in period had 400 instructors agree to participate in the study. Demographic breakdown of the instructors is seen in Table 1. Two-thirds of the instructors are female, while almost three-fourths of the instructors are white. The 400 instructors teach 849 classes with 24,861 unique students in them. Counting students who take more than one class, there are 33,975 total student observations. Size of the classes widely varies. The COVID-19 Pandemic made many courses online, subsequently pushing introductory courses to combine sessions and contain as many as 1,100 students. Conversely, there are roughly 9,000 students in classes smaller than 40 people with the largest percent of students (17%) in classes between the sizes of 20-29 people.

Though there are over 33,000 students in the study, we expect the final number to be less. Student evaluation completion rates consistently hover around 60% of total students in the class. Moreover, technology constraints inhibit us from implementing the treatment scripts through the mobile application. With roughly one-third of all student evaluations completed through the mobile application in recent semesters, we expect our final dataset to contain around 14,000 individuals. However, because we do not know for certain who will complete the survey or not, we cannot be sure what are final dataset will contain.

At the end of the Spring 2021 semester (early May), we will combine each participating student's: (1) gender; (2) race; (3) year/rank; and (4) major with their (5) end-of-course official grade and (6) treatment. We also will utilize data on the participating instructors' (1) gender; (2) race/ethnicity, (3) age; (4) department; (5) track; (6) years in track; (7) years at Ohio State and (8) full-time or part-time faculty status.

We expect to complete our analysis over the summer of 2021. Instructors and their unit heads will be given expanded SEI reports with scores disaggregated by treatment and control groups, as well as guidelines for how to incorporate this information in performance evaluation.

Analysis

<u>Statistical Methods:</u> The use of a randomized control trial allows us to circumvent issues of endogeneity or identification issues. For these reasons, we utilize OLS regressions to measure changes in the average score. We also utilize a linear probability model to estimate the probability of completing the evaluations as a function of treatment assignment. Missing data will be considered on a case-by-case basis, but primarily, will be analyzed as long as they have completed the student evaluations of instructors. Outliers will be determined by the interquartile range method. Regressions will be run with and without outliers to compare how different the results are.

<u>Statistical Models:</u>

The OLS Model:

$$\begin{aligned} EvaluationScore_i = \alpha_0 + \beta_1 Gender_s + \beta_2 Rank_s + \beta_3 Major_s + \beta_4 Race_s \\ + \beta_5 CourseGrade_s + \beta_6 Treatment_s + \eta_c + \epsilon_s \end{aligned}$$

where i represents each teacher, s represents each student, and η_c represents class-level fixed effects.

The Linear Probaility Model:

$$\begin{aligned} COMPLETE_s = \alpha_0 + \alpha_1 Gender_s + \alpha_2 Rank_s + \alpha_3 Major_s + \alpha_4 Race_s \\ + \alpha_5 CourseGrade_s + \alpha_6 Treatment_s + \eta_c + \epsilon_s \end{aligned}$$

where $COMPLETE_s$ is a $\{0, 1\}$ binary variable in which values of 1 indicates the student completed their SEI and 0 if they did not, s represents each student, and η_c represents class-level fixed effects.

<u>Robustness Checks:</u> To measure the impact manually changing randomization has, as laid out above, we will run regression analysis keeping only the true randomization values. Specifically, any treatments that were manually changed for consistency will be dropped in order to measure the impact the manual changes had.

<u>Heterogeneous Effects:</u> We plan to evaluate the differential impact of treatments on instructor subgroups. Our main analysis will examine the differences between male and female instructors but will also extend into comparing races and ethnicities. We will further break-out these categories into age groups, departments, tenure/non-tenure tracks, and time at the university. A generic table is presented below for treatment analysis, by gender. We will also analyze the presence of “ethnic affinity," where students of the same ethnicity rate

the instructor higher than their counterparts. We plan to also examine heterogeneity by student race and gender as previous results by Peterson et al. (2019) were driven almost exclusively by white male students.

ADDITIONAL ANALYSIS: Additional analysis will look at the interaction between course grade and treatments. Particularly, if there is a strong correlation between a student's grade and how they evaluate the course, is this correlation reduced or enhanced by the treatment. This analysis will be run with and without course fixed effects as well as further analysis with student fixed effects to look at the students enrolled in multiple classes.

Because we have data on when students have completed their SEIs, we will also be able to analyze dosage/decay effects for students. Specifically, we will be able to analyze if the treatment effects decay for students in multiple classes as they complete their SEIs. Placebo tests will be measured for students who were assigned to treatment groups but completed the SEI via the mobile app, thus not seeing any treatment. To avoid multiple hypothesis testing concerns, we will utilize PCA indices to group like questions together. Question 10 will always be included separately and never in an index as it is the main instrument of our analysis.

Interpretation

To assess the impact of the study on the instructor's SEIs, we will compare the instructor's scores for the control group to the scores for the treatment group(s).

- Comparison to the "high stakes" treatment group indicate the impact of showing students text noting the importance of SEIs in faculty performance evaluation, tenure, and promotion.
- Comparison to the "implicit bias" treatment group indicate the impact of showing students text noting the role of implicit bias in subjective evaluations and reminding students to focus on aspects of course instruction distinct from characteristics of the instructor.
- Comparison to the "combined" treatment group indicate the impact of showing students both sets of text described above.

Comparison of unit/college/University average scores for the control group only to unit/college/University average scores for those not participating in the study should be interpreted as the average impact of participating in the study without being exposed to any new introductory language before completing the SEI. These should be considered as spillover effects of how simply participating in the study, without directly receiving treatment, affects student responses.

Comparison of the instructor's scores for the control group only to unit/college/University average scores for those not participating in the study should be interpreted as the individual-specific impact of participating in the study. This should be taken into account to the extent that participation in the study has heterogeneous effects for instructors at higher risk of facing implicit bias.

Appendix

Table 1: Participating Instructors

	Total	**Male**	**Female**
	Gender		
Overall	398	134	264
	Race & Ethnicity		
American Indian	1	1	0
Asian	37	14	23
Black	13	4	9
Hispanic	23	10	13
Two or More Races	7	5	2
Undisclosed	24	6	18
White	293	94	199
	Age		
19-28 Years	70	23	47
29-38 Years	110	36	74
39-48 Years	109	30	79
49-58 Years	61	21	40
59-68 Years	39	17	22
69-78 Years	9	7	2

Notes: Demographic data is available for 398 out of 400 instructors. Two instructors are currently missing data.

Table 2: Comparison of Instructor Demographic Means

	Opt-In	**Opt-Out**	**T-Statistic**
Gender and Age			
Female	66.3%	42.5%	8.87***
Age	41.3	40.7	0.75
Race & Ethnicity			
American Indian	0.2%	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	1.7%	2.3%	0.68
Undisclosed	6.0%	7.3%	0.94
White	73.6%	67.1%	2.54**
FTE			
FTE of 0.25	3.5%	3.9%	0.40
FTE of 0.50	25.1%	38.0%	4.94***
FTE of 0.75	2.3%	3.5%	1.27
FTE of 1.00	55.8%	40.5%	5.66***
Job Title			
Assistant Professor	10.8%	4.8%	4.78***
Associate Professor	17.8%	11.2%	3.71***
Professor	14.6%	12.3%	1.22
Graduate Teaching Associate	20.9%	31.7%	4.36***
Lecturer	13.1%	17.3%	2.08**
Senior Lecturer	8.8%	5.9%	2.20**
Tenure			
Tenure Track	42.7%	29.0%	5.46***
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***

Notes: Reported T-Statistics are calculated using a two sample t-test comparing group proportions and means. Significance of t-scores are 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. “Opt-in" includes all instructors who agreed to participate in the study. “Opt-out" includes all instructors who did not consent to the study but were eligible to participate.

Table 3: OLS Results - Gender

	Implicit Bias		**High Stakes**		**Combined**	
	Male	*Female*	*Male*	*Female*	*Male*	*Female*
Overall						
Age						
19-28 Years						
29-38 Years						
39-48 Years						
49-58 Years						
59-68 Years						
69-78 Years						
Classification						
Student						
Faculty						
Track						
Tenured						
Non-Tenured						
Position						
Lecturer						
Senior Lecturer						
Assistant Professor						
Associate Professor						
Professor						

Notes: Reported values are interpreted as the change in scores in relation to the control group. Positive values indicate the treatment group gave evaluations that are higher relative to the control group. Negative values indicate the opposite.

References

[1] Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research.*

[2] Chávez, K., & Mitchell, K. M. (2020). Exploring bias in student evaluations: Gender, race, and ethnicity. *PS: Political Science & Politics*, 53(2), 270-274.

[3] Cummings, R.G. and L.O. Taylor, 1999, "Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method," *American Economic Review* 89, 649 – 665.

[4] Holman, M., Key, E., & Kreitzer, R. (2019). Evidence of bias in standard evaluations of teaching. *http://www.rebeccakreitzer.com/bias/*

[5] MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4), 291-303.

[6] Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, 17(2), 535-566.

[7] Peterson, D. A., Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. *PloS one*, 14(5), e0216241.