Research Protocol

I. Objectives

The purpose of this study is to assess the efficacy of utilizing modified introductory language to mitigate implicit bias in student evaluations of instruction. So-called "cheap talk" scripts will presented to survey respondents in advance and describe the hypothetical biases that tend to arise in the study context (Cummings and Taylor, 1999). 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 under-represented populations.

II. Background and Rationale

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 (see, e.g., Boring, Ottoboni, and Stark 2016; Mengel, Sauermann, and Zolitz 2018; Holman and Kreitzer, 2019). This holds 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 Forthcoming). 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, 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 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.

III. Procedures

a. Research Design

This research will utilize a "survey experiment" design, which involves embedding randomized text within a survey given to subjects. While all subjects will see the same questions (the current questions asked on the OSU Student Evaluation of Instruction), the introductory paragraph will vary across respondents.

b. Sample

To be able to recover the effects on the order of magnitude reported in Peterson et al. (2019) with 80% power, we calculate that we will need 1,500 responses in each treatment arm, for a total of 6,000 responses. We will recruit faculty who are teaching undergraduate courses in the Spring 2021 term in the

Colleges of Arts and Sciences and in the College of Food, Agricultural, and Environmental Sciences. Because our focus is on bias against gender and racial minorities, we will particularly focus on ensuring recruitment of instructors of diverse backgrounds. Our plan is to send e-mail invitations to all instructors in these two colleges describing the study and including an embedded link they can use to opt-in for the study. Additionally, we will work with the Office of Diversity and Inclusion to send an additional invitation to under-represented minority faculty. Instructors who opt-in will have students in their Spring 2021 undergraduate courses randomly assigned to one of the four treatment arms for the end-of-course evaluations. Note that since a single student may be enrolled in more than one class that is included in the experiment, the total number of students is likely to be less than the 6,000 total number of responses. We will also ensure that students are assigned to the same treatment arm across all of the courses they are enrolled in this semester that opt into the experiment.

c. Measurement/Instrumentation

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.

d. Detailed Study Procedures

Once an instructor opts into the study, we will send a notification e-mail to the relevant department chair notifying him or her of the instructor's participation and noting that the SEI results for this semester may be impacted. We will then download the course rosters from SIS for the sections taught by instructors who opt in to the study and extract a list of unique "Emplids" (student numeric identifiers) across all of the sections. Each unique Emplid will be assigned to one of three treatment arms: high stakes, bias, and combined.

Within each course, students will be randomly assigned to treatment and control groups. However, across courses, assignment into treatment groups will be dependent on class size, unit, and course and instructor characteristics in order to maintain statistical balance and to optimize reporting of disaggregated SEI results. For example, in smaller courses, students may be split only between the control group and one treatment group, so that responses within each group remain large enough for instructors to receive disaggregated SEI scores. Similarly, depending on the race/gender composition of participating instructors within units, allocation to treatment arms may be adjusted by study personnel to ensure balance based on instructors' demographic characteristics.

After the randomization has been done, the PIs will send a listing of Emplids, associated course codes, and the treatment arm to which each student has been assigned to the university's SEI administrator, who

will program the backend to ensure that each student sees the right version of the SEI instrument for each course that is participating in the pilot.

At the end of the semester, the PIs will extract from SIS each student's: (1) gender; (2) race; (3) year/rank; (4) major; and (5) end-of-course official grade. The SEI administrator will provide us a file each student's SEI responses for each course that is participating. We will merge this file with the SIS data using the Empid variable, and will then delete this variable from the final analysis dataset. So the final dataset will contain only: (1) student gender; (2) student race; (3) student year/rank; (4) student major; (6) course ID; (7) end-of-course official grade; (8) SEI responses for that course.

PIs will also extract information from OSU HR on instructors' race and gender to assess whether treatment effects differ for women and persons of color.

Aside from the Emplid variables, no individually-identifiable information will be saved by the research team, and the Emplids will be removed after the final matching step, so no unique student identifiers will remain in the final analysis dataset.

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.

e. Internal Validity

Because students will be randomly assigned to treatment (and treatment will be consistent across courses, avoiding potential spillovers), we expect the proposed design to have high internal validity. Since we are also using the existing SEI instrument and platform (so the student experience will be identical to their usual evaluation process), we also expect the study to have very high external validity.

It is expected that the intervention will increase SEI scores, on average, particularly for instructors in historically under-represented groups. A previous study conducted by researchers at Iowa State University and published in PLOS One, 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.

It is, however, possible that the intervention could reduce SEI scores, if the treatment scripts generate animosity towards instructors from under-represented groups. Randomization within courses limits the risks associated with this, as it ensures that no groups of instructors or courses will be systematically disadvantaged. Moreover, the expanded SEI reports will allow instructors, as well as their unit heads, to assess SEI scores among the control group, which receives the usual set of instructions.

There is, however, still some possibility that control groups may be contaminated by the presence of treated groups within the same courses/units. In this case, SEI scores for participating instructors in majority groups may be used as a comparison group. In order to produce relevant comparison scores, we will obtain data from the Registrar's office on average SEI scores for instructors not participating in the intervention, disaggregated by class size and college, as follows. These comparisons will help mitigate any potential negative effect on SEIs due to participation in the study. The guidance below will also be included in the expanded SEI report to help instructors and unit heads interpret SEI scores for those participating in the study.

To assess the impact of the study on the instructor's SEIs, 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.

To assess the instructor's SEIs relative to other instructors, compare the instructor's scores to unit/college/University average scores within control/treatment groups.

- Comparisons within the control group should be interpreted as responses in the absence of priming regarding implicit bias and the high stakes associated with SEIs. That is, participating in the study but not receiving any treatment.
- Comparisons within the "high stakes" treatment group should be interpreted as responses when students are shown text noting the importance of SEIs in faculty performance evaluation, tenure, and promotion.
- Comparisons within the "implicit" treatment group should be interpreted as responses when students are shown 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.
- Comparisons within the "combined" treatment group should be interpreted as responses when students are shown 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. This should be taken into account to the extent that 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.

d. Data Analysis

Given the randomization into treatment and control groups within courses, estimation of treatment effects is straightforward and will be done using sub-group analysis and multiple regression to account for differences across units and across course type (general/specialized, large/small, time of day, etc.) and student type (gender, race, major, rank, etc.). We will focus on questions relating to the overall evaluation of the instructor but also examine how the effect of the intervention differs across various types of questions in the course evaluation instrument.

The particular circumstances around the COVID-19 pandemic make remote and virtual instruction an increased likelihood for all classes. Given this likelihood, results may differ from previous studies that primarily used in-person classes. The current impact of the mode of teaching is ambiguous. While Rovai et al. (2006) found a negative impact on instructors' scores for online courses, Kelly et al. (2007) reported no statistical differences between the modes of teaching. Moreover, while previous comparative studies also dealt with sorting concerns, the COVID-19 pandemic may assuage those concerns because of the

widespread and systemic use of virtual learning. In our analysis, we will clearly note the differences in virtual and in-person classes, as well as synchronous and asynchronous learning.

IV. Bibliography

Boring, Anne, Kellie Ottoboni, and Philip B. Stark, 2016, "Student evaluations of teaching (mostly) do not measure teaching effectiveness," ScienceOpen Research.

Chavez, Kerry, and Kristina M.W. Mitchell, Forthcoming, "Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity," PS: Political Science and Politics.

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.

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

Kelly, H. F., Ponton, M. K., & Rovai, A. P. (2007). A comparison of student evaluations of teaching between online and face-to-face courses. The Internet and higher education, 10(2), 89-101.

Mengel, Friederike, Jan Sauermann, and Ulf Zolitz, 2019, "Gender Bias in Teaching Evaluations," Journal of the European Economic Association 17(2): pp. 535-566.

Peterson, David A.M., Lori A. Biederman, David Andersen, Tessa M. Diitonto, and Kvin Roe, 2019, "Mitigating gender bias in student evaluations of teaching." PLOSOne

Rovai, A. P., Ponton, M. K., Derrick, M. G., & Davis, J. M. (2006). Student evaluation of teaching in the virtual and traditional classrooms: A comparative analysis. The Internet and Higher Education, 9(1), 23-35.