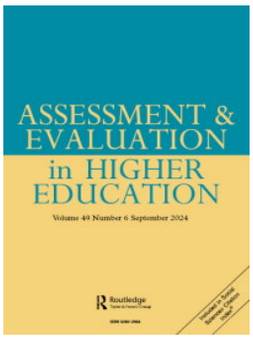

Assessment & Evaluation in Higher Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/caeh20

Can you mitigate gender bias in student evaluations of teaching? Evaluating alternative methods of soliciting feedback

Ann L. Owen, Erica De Bruin & Stephen Wu

To cite this article: Ann L. Owen, Erica De Bruin & Stephen Wu (02 Oct 2024): Can you mitigate gender bias in student evaluations of teaching? Evaluating alternative methods of soliciting feedback, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2024.2407927

To link to this article: https://doi.org/10.1080/02602938.2024.2407927

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2024.2407927

Routledge
Taylor & Francis Group

Can you mitigate gender bias in student evaluations of teaching? Evaluating alternative methods of soliciting feedback

Ann L. Owen[a], Erica De Bruin[b] and Stephen Wu[a]

[a]Economics Department, Hamilton College, Clinton, NY, USA; [b]Government Department, Hamilton College, Clinton, NY, USA

ABSTRACT

Student evaluations of teaching can be unreliable indicators of effective teaching and affected by implicit bias. We conduct a randomized experiment at a selective U.S. liberal arts college in which we vary both the instrument and timing at which we solicit student feedback to assess whether either intervention can mitigate gender disparities in qualitative evaluation comments. External evaluators with expertise in teaching evaluation scored the positivity, specificity, and constructiveness of student comments in control and treatment groups. We find neither intervention is successful in mitigating gender disparities. Students that receive an alternate prompt which highlights the potential for implicit bias and asks them to articulate and apply their own criteria for effective teaching penalize female faculty at similar rates as those receiving a more standard evaluation prompt. We also find no evidence that delaying solicitation of student feedback until the start of the next semester, when students face less stress and fatigue, reduces the gender gap in the positivity of feedback. Instead, we find substantial evidence that the positivity of qualitative student evaluations is affected by student, instructor, and course characteristics across all study conditions. The results suggest that bias in student evaluations may be difficult to mitigate.

KEYWORDS

Teaching evaluations; gender bias; university policy

Introduction

The evaluation of teaching plays a crucial role in the career paths of college faculty. Decisions about hiring, tenure, promotion, and salary can all be significantly impacted by the assessment of a faculty member's teaching. Although there are many ways to assess teaching effectiveness, including self-evaluation through personal statements, peer review, and pre-post learning assessments, student evaluations of teaching (SETs) play an outsized role at many institutions. However, there is ample research showing that SETs can be unreliable indicators of effective teaching (e.g. Carrell and West 2010; Boring, Ottoboni, and Stark 2016; Uttl, White, and Gonzalez 2017). Student evaluations are also prone to implicit bias, with the strongest evidence showing that female faculty and others underrepresented in their fields receive lower numeric scores and more negative comments about their personality, appearance, competence, and professionalism, among other issues (MacNell, Driscoll, and Hunt 2015; Boring 2017; Mengel, Sauermann, and Zölitz 2017; Aragón, Pietri, and Powell 2023).

CONTACT Ann Owen aowen@hamilton.edu Economics Department, Hamilton College, Clinton, NY 13323, USA.

The consequences of using such an unreliable measure to evaluate faculty teaching are stark: colleges and universities make important decisions that affect who becomes and stays a college professor based on flawed evidence. Student evaluations can create perverse incentives for faculty to teach in ways that may detract from deep learning and contribute to grade inflation (Stroebe 2020). Receiving cruel or disrespectful comments, which is more common for faculty from underrepresented groups, can also cause faculty to divert time from research and take a toll on mental health (Lindahl and Unger 2010). The problems with SETs, and with qualitative comments in particular, have led some scholars to recommend restricting or eliminating their use altogether (e.g. Kreitzer and Sweet-Cushman 2022). Yet providing students with the opportunity to provide feedback on teaching is important. Despite their limitations, qualitative comments have the potential to provide unique insight into learning and teaching practices from a student perspective (Alhija and Fresko 2009; Appleton 2018; Steyn, Davies, and Sambo 2018).

While there are many studies that document the bias in SETs and their failure to reflect effective teaching, few provide evidence-based guidance about what to do about it. Some recent scholarship has sought to address this gap. For quantitative scores, reducing the size of the scale has been found to mitigate gender bias (Rivera and Tilcsik 2019). Other recent studies test whether informing students about the potential for implicit bias can reduce it, with mixed results. For example, in a randomized controlled trial, Peterson et al. (2019) find that informing students about gender biases had a positive effect on quantitative evaluation scores for female faculty; however, other studies have not replicated this finding in other settings (e.g. Key and Ardoin 2019a, 2019b). Meanwhile, other work has found that providing information on past discrimination reduced biases against female faculty, but normative statements reminding students not to discriminate did not affect scores (Boring and Philippe 2021).

Yet there remains much we do not know about potential interventions to reduce bias in student evaluations of teaching. Existing studies have focused primarily on informational treatments—informing students about the existence of bias or the potential for it, without evaluating whether other aspects of the process or instrument through which student feedback is solicited affects student responses. Moreover, to our knowledge, no studies have tested interventions to mitigate bias in qualitative comments, where it may be most acute (Mitchell and Martin 2018; Wallace, Lewis, and Allen 2019).

To address this need, we conduct an experiment at a residential liberal arts college in the United States in which we vary both the method and timing at which we solicit student feedback. Forty tenured faculty members teaching more than 200 courses across the curriculum volunteered to participate. Building on existing experiments, we combine an informational treatment, which informs students about implicit bias and how it can be mitigated, with changes to the instrument used to solicit feedback and the timing at which feedback is solicited. These interventions are designed to minimize the conditions associated most strongly with implicit bias. The aim is to evaluate whether either intervention can reduce one of the most well-documented forms of implicit bias in evaluations: bias against female faculty in qualitative feedback.

The experiment randomly assigned students within each course taught by faculty members in the study into three groups: a control group and two treatment groups. The control group completed a standard evaluation form, comparable to those at many institutions, which includes a mix of directed qualitative and quantitative questions, at the usual end-of-semester time period. Students in the two treatment groups completed an alternate prompt that encouraged a more reflective and open-ended response, and informed students about the potential for implicit bias to affect evaluation. Those in the first treatment group completed their evaluations at the end of the semester, while those in the second completed them at the beginning of the following semester. This design allows us to evaluate the impact of both the nature of the prompt as well as the timing of the solicitation. We then asked external readers, selected for their expertise in the evaluation of teaching, to rate how positive the qualitative feedback provided in each

individual evaluation was, along with how specific and constructive it was and whether it contained explicit references to identity-based attributes of the instructor.

The results suggest that neither intervention is a panacea: We find no reduction in the penalty for female faculty when varying the timing or method through which student input is sought. Instead, we find strong evidence that across all study groups, student, instructor, and course characteristics affect student evaluations. In particular, female faculty, faculty that award lower grades relative to other faculty, and faculty teaching larger courses all receive less positive evaluations from students. We also find evidence of interactions between instructor and course characteristics that widen the positivity gap between male and female faculty. For example, female instructors pay a larger penalty for harsh grading and large classes, across all study groups. Neither the alternative prompt nor the alternative prompt combined with delayed timing resulted in more specific or constructive feedback, though external evaluators rated student responses provided at the start of the subsequent semester more positively for all faculty. While more research into interventions is needed, our findings suggest that mitigating measurement and equity bias in student evaluations may be difficult to do through changes in question wording or the timing at which feedback is solicited.

Experimental design

We use a randomized controlled trial to test two interventions designed to reduce bias and improve the usefulness of feedback in qualitative student evaluation comments. The first intervention varies the instrument that solicits feedback from students. The new instrument: (1) informs students of the potential for implicit bias in the evaluation of teaching, and how they can reduce it; (2) directs students to identify the specific criteria they consider in judging teaching effectiveness; and (3) asks them to evaluate the course and instructor in relation to those criteria. The second intervention delays the timing at which feedback is solicited from the traditional end-of-semester period to the start of the *subsequent* semester.

These interventions are motivated by existing scholarship on the conditions under which implicit bias is most likely to occur. Past research has shown that people are more likely to resort to stereotypes in ambiguous situations (Snyder et al. 1979; Katz 2014). It also emphasizes that implicit bias is common under time pressure, when people face fatigue or cognitive overload, and when people lack solid information to make a decision (Aronson 2008; Johnson et al. 2011; Ma et al. 2013). There is also evidence that informing people about implicit bias can help reduce it (Nadler et al. 2014; Axt, Casola, and Nosek 2019), although as we note above, findings are mixed in the context of student evaluations specifically. Other studies suggest that providing students a clear structure for their qualitative feedback, as opposed to free text entry, results in more constructive responses (Reja et al. 2003; Hoon et al. 2015).

The alternate prompt thus encourages students to explicitly articulate the concrete features of teaching and the classroom experience that matter most for their own learning, and then to evaluate faculty against these criteria—reducing ambiguity and providing students a way to structure their responses. Delaying the solicitation of student feedback enables students to provide their evaluations when they have fewer demands on their time and are less fatigued, and thus less likely to lean on stereotypes. We hypothesize that both interventions will reduce the gender gap between male and female faculty in how positive student evaluations are (H1).

In designing the alternate instrument, we also sought input from students, faculty, and administrators at our institution. We held a series of meetings between September 2020 and May 2021 to identify the aspects of the existing system of evaluations each group found most valuable, and aspects that limited their usefulness or discouraged students from responding thoughtfully. Student input was especially useful as several reported not being able to spend much time providing feedback at the end of the semester when final papers and examinations are their priority. Some also expressed frustration that the questions on the current survey were repetitive and

would prefer an evaluation that allowed them to tailor their own responses to what they thought was most important.

The experiment included 210 courses taught by 40 tenured faculty members at a selective U.S. liberal arts college in Fall 2022 and Spring 2023. Prior to conducting the experiment, we received approval from our institution's Institutional Review Board (IRB Project S22-045.) All faculty and students who participated provided informed consent. We invited all tenured faculty at this institution to participate in the study. The included courses represent a little over 20% of all courses that were taught that academic year. At this institution, SETs are conducted online. Students receive an email with the link to evaluate the instructor for each course and normally can complete the evaluations at any time in the two weeks preceding the beginning of the final examination period.

For each course, we randomly assigned all students to one of the three study groups. The first group of students (group 1) served as the control group and received the standard prompt, consisting of three qualitative prompts pertaining to the course and three pertaining to the instructor, during the usual time period at the end of the semester. Students also rated the course and instructor along several dimensions on a 5-point scale. In the treatment conditions, students received an alternate prompt either at the end of the semester (group 2) or in the second and third weeks of the following semester (group 3). One notable difference for the third study group relative to the other two groups relates to the incentives to respond. For groups 1 and 2, end-of-semester grades would be withheld for an extra two weeks if students failed to complete their teaching evaluations, whereas this external incentive was not present for group 3. The specific wording of the prompts for the control and treatment conditions can be found in Table 1.

To assess the nature of qualitative student feedback across the three study groups, we hired three faculty members from peer institutions who have expertise in the scholarship of pedagogy, student learning, and faculty assessment. We asked each person to independently score every individual student evaluation on three dimensions using a 5-point scale: how detailed or specific the evaluation was, how constructive in providing guidance to improve, and how positive or negative it was overall. To assess the impacts of alternative methods of soliciting feedback on the gender gap, we primarily focus on the positivity of student comments, but we also examine differences in specificity and constructiveness to gain further insight. The presentation of each evaluation to the readers was random and anonymous *via* an online viewing system. Scores were highly correlated across the three readers and we averaged the three scores on each evaluation in our analysis. The questions the readers responded to in scoring individual student evaluations are shown in Table 2.

Table 3 provides summary statistics for all the data used in our analysis, first for all groups, and then within each study group. As Table 3 shows, the response rate was significantly lower for group 3 than for the other groups, but there is little difference in responses between the first two groups. The sample sizes for groups 1, 2, and 3 are 1,042, 994, and 678, respectively, which corresponds to response rates of 87%, 89%, and 62%. In terms of the external evaluators' ratings of student feedback, group 3 was rated on average to be slightly more positive (4.52 out of 5) than groups 1 (4.43) and 2 (4.40), while evaluations in group 1 were rated as more specific (3.15) and constructive (2.81) than group 2 (3.06 and 2.66 for specificity and constructiveness, respectively) or group 3 (2.97 and 2.59 for specificity and constructiveness, respectively). Although students in group 1 were asked three separate questions about their instructor (compared to only one question for groups 2 and 3), they still wrote the fewest words out of all groups (93 words combined across all three instructor-focused questions). The length of the evaluations of the instructor were somewhat shorter in group 3 (101 words on average) than in group 2 (114 words). Table 3 also shows the means for student and faculty characteristics used in the analysis. Appendix Table A1 shows the results of balance tests which find a standardized difference in means of less than 0.10 for all covariates, validating the random assignment of students to each study group.

Table 1. Qualitative student evaluation questions.

Study group	Questions
Group 1	1. Describe your effort and engagement in this course. 2. Describe the learning environment in this course. 3. Please assess the overall quality of the course and explain your reasoning. 4. Describe the instructor's interactions with students. 5. Describe the instructor's responses to your work. 6. Please assess the overall effectiveness of the instructor and explain your reasoning.
Groups 2 & 3	First, take a moment to reflect on what an effective teacher does. Effective teaching is multifaceted and we encourage you to think broadly. Research has shown that implicit bias can affect the evaluation of faculty. You can reduce the impact of implicit bias in your evaluation by clearly stating the criteria you are using and reflecting on whether the criteria or your application of it is influenced by irrelevant aspects of the instructor's identity such as race, ethnicity, gender, age, or sexuality. 1. Please list briefly the criteria you consider in judging the effectiveness of a faculty member at [institution]. What qualities are most helpful for your learning? 2. Please share your learning experiences and the classroom environment in the course by telling us, for example, how the course challenged you to think critically and creatively, the effectiveness of the assignments and assessments in furthering your learning, the quality of feedback you received, the interactions with the faculty that you had outside of the classroom, and about the inclusivity of your learning experience. Connect your comments to your understanding of effective teaching that you articulated above.

Table 2. Evaluator questions for scoring individual student evaluations.

To what extent does the evaluation provide feedback that is specific and detailed in its descriptions of the course and instructor, as opposed to more general or vague?
To what extent does the evaluation provide feedback that is constructive in providing guidance that would enable the instructor receiving it to improve their teaching?
Overall, how positive or negative was the student's feedback on the course and/or instructor?

Evaluators were also asked to identify evaluations that contained explicit bias, but fewer than 1% of all evaluations were identified as such and we do not use those results in our analysis.

Results

We now turn to the analysis of how various student, course, and faculty characteristics impact three outcomes: the positivity, specificity, and constructiveness of student feedback, as measured by the average ratings of three independent external evaluators. We begin by noting that there are strong relationships among these characteristics of feedback. Unsurprisingly, there is a strong positive relationship between specificity and constructiveness, with a correlation coefficient equal to 0.73. Even with the high correlation between specificity and constructiveness, it is notable that there are more evaluations rated as specific than constructive. In our data, evaluations rated as constructive are almost always specific, but evaluations rated as specific are not always constructive. Meanwhile, evaluations that are more positive in tone are less likely to be specific (correlation = −0.13) or constructive (correlation = −0.39), suggesting that students are more specific or constructive when they have something negative to say, which is consistent with research that shows that the length and helpfulness of online reviews is negatively correlated with consumer sentiment (Eslami, Ghasemaghaei, and Hassanein 2018; Ghasemaghaei et al. 2018). When separating the sample by study group, the correlation coefficients are remarkably similar across all three groups.

Consistent with the negative correlation between specificity and positivity, students write fewer words in evaluating an instructor when the evaluation was rated more positively by the outside evaluators. As is clear from the descriptive statistics in Table 3, overall, the evaluations are positive, and 56% of them were rated a 5 for positivity by all three outside reviewers. Those positive evaluations have, on average, 93 words evaluating the instructor (with averages of 88 for study group 1, 99 for study group 2, and 92 for study group 3). In contrast, the evaluations that were rated as being less positive (average scores of between 2 and 4), have an average number

Table 3. Summary statistics.

	All Groups				Study Group 1				Study Group 2				Study Group 3			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Variables	mean	sd	min	max	mean	Sd	min	max	mean	sd	min	max	mean	sd	min	max
Evaluator Ratings																
Positivity	4.441	0.847	1	5	4.431	0.792	1	5	4.400	0.909	1	5	4.516	0.828	1	5
Specificity	3.072	1.108	1	5	3.154	1.115	1	5	3.056	1.095	1	5	2.970	1.111	1	5
Constructiveness	2.703	0.755	1	5	2.811	0.763	1	5	2.665	0.757	1	5	2.593	0.719	1	5
Student Characteristics																
Pell Grant Recipient	0.196	0.397	0	1	0.210	0.408	0	1	0.188	0.391	0	1	0.184	0.388	0	1
White Student	0.658	0.474	0	1	0.664	0.473	0	1	0.659	0.474	0	1	0.647	0.478	0	1
Female	0.593	0.491	0	1	0.599	0.490	0	1	0.581	0.494	0	1	0.600	0.490	0	1
Course Grade/Term GPA	0.974	0.143	0	2.281	0.973	0.145	0	1.574	0.976	0.145	0	2.281	0.971	0.139	0	1.600
Faculty Characteristics																
Female Faculty	0.503	0.500	0	1	0.507	0.500	0	1	0.504	0.500	0	1	0.494	0.500	0	1
5–10 years experience	0.300	0.458	0	1	0.300	0.459	0	1	0.298	0.458	0	1	0.301	0.459	0	1
10–20 years experience	0.119	0.324	0	1	0.118	0.323	0	1	0.116	0.320	0	1	0.125	0.331	0	1
20+ years experience	0.569	0.495	0	1	0.567	0.496	0	1	0.571	0.495	0	1	0.569	0.496	0	1
Avg (Grade/Term GPA) all students	0.969	0.061	0.817	1.101	0.969	0.061	0.817	1.101	0.968	0.061	0.817	1.101	0.968	0.062	0.817	1.101
Class Size	24.12	12.46	4	90	23.85	12.56	4	90	23.90	12.42	4	90	24.85	12.33	4	90
Observations	2,714				1,042				994				678			
Response Rate	0.80				0.87				0.89				0.61			
Avg Words Evaluating Instructor	103				93				114				101			

Evaluator Ratings are the average of three evaluators for each evaluation. Avg Words Evaluating Instructor is the average of the total of questions 4, 5 and 6 for Study Group 1 and Question 2 for Study Groups 2 and 3.

of words evaluating the instructor of 101 words. The most negative evaluations (rated as a 1 for positivity by all three evaluators) have an average of 128 words.

The most negative evaluations comment on a range of issues, including grading standards ('graded tests and homework's extremely harsh'), class environment ('cold' or 'stiff'), or lack of learning ('I don't think I could tell you anything I really learned in this course'). The comments on the very positive evaluations cover some of the same themes but also are more likely to make positive comments about the personal attributes of the instructor. Some commented on the expertise of the instructor ('Professor X had a very strong understanding of the material'), the enthusiasm of the professor ('the professor was very enthusiastic about the material which made learning more fun'), or simply declared the professor to be 'fantastic' or 'the best I've had.' As was the case with the negative evaluations, students wrote about how much they learned ('helped me grow as a writer') and class environment ('learning environment was positive and collaborative'). In the positive evaluations, comments about grading standards take a different tone ('tough but fair').

Frequency analysis of commonly used words reveals some differences across study groups. In particular, while students in all three study groups commented on availability and helpfulness of the faculty member outside of class, that type of comment was more common for those in study group 1, who were prompted specifically to discuss the professor's interactions with students. Similarly, students in all groups often discussed the level of challenge a course presented, but students in study groups 2 and 3 who responded to a more open-ended prompt were more likely to mention the extent to which they were challenged to think critically or creatively. Unfortunately, our sample was not large enough to perform a robust analysis of gender differences in the frequency of specific words across different study groups.

Regression results in Table 4 examine more systematically how the positivity of student evaluations varies by faculty, student, and course characteristics, as well as by the method of soliciting feedback. The dependent variable is the average positivity rating from three external evaluators. Model 1 shows the results of regressing the positivity rating on study group, class size, faculty gender, and interactions between the study group and gender. Model 2 includes an interaction between faculty gender and the 'harshness' of their grading, relatively to other faculty. To capture the harshness of faculty grading, we calculate, for each student in a course, the ratio between their course grade and overall term GPA, and then average that across all students in a given course. Model 3 includes other student and faculty characteristics, including whether the student is a Pell grant recipient, the student race and gender, how the student's grade in the course compares to their own overall GPA, and years of teaching experience for the instructor. Pell grants are need-based grants given to low-income students and we use receipt of a Pell grant as an indicator of socio-economic status. Finally, Model 4 includes interactions between faculty gender and study group, grading, and course size. Existing scholarship has found both class size and grading to be strong predictors of evaluations (McPherson 2006; Bedard and Kuhn 2008; Miles and House 2015; Berezvai, Lukáts, and Molontay 2021).

Figure 1 illustrates these effects, showing predicted probabilities associated with the specification in Column 4 of Table 4. In this graph, the top panel compares the predicted positivity of qualitative comments from students about male and female faculty, by study group. Error bands of two standard errors in either direction are shown. The middle and bottom panels compare male and female faculty that are harsh graders and that teach large classes, respectively. To identify faculty that are harsh graders, we calculate for each student taught by a faculty member the grade that student earned in the class divided by the GPA the student earned that term and average that statistic across all the students a faculty member taught in a given term. Low grading faculty have averages that are two standard deviations below the average. We count faculty teaching classes that are two standard deviations above the average class size as faculty with large courses.

Table 4. Regression results for positivity.

	(1)	(2)	(3)	(4)
Study Group 2	−0.063	−0.059	−0.025	−0.061
	(0.051)	(0.051)	(0.041)	(0.050)
Study Group 3	0.100*	0.102*	0.092**	0.097*
	(0.053)	(0.053)	(0.043)	(0.053)
Class Size	−0.005***	−0.005***	−0.004***	−0.001
	(0.001)	(0.001)	(0.001)	(0.002)
Female Faculty	−0.189***	−1.162*	−0.183***	−0.941
	(0.050)	(0.604)	(0.038)	(0.606)
Study Group 2*Female Faculty	0.055	0.054		0.056
	(0.074)	(0.075)		(0.075)
Study Group 3*Female Faculty	−0.022	−0.020		−0.014
	(0.080)	(0.080)		(0.080)
Avg (Grade/Term GPA) all students		0.774*	0.929**	0.838**
		(0.415)	(0.400)	(0.416)
Female Faculty*Avg(Grade/Term GPA) all students		1.001		0.940
		(0.619)		(0.618)
Female Faculty*Class Size				−0.007***
				(0.003)
Pell Grant Recipient			−0.005	
			(0.051)	
White Student			−0.038	
			(0.042)	
Female Student			−0.105***	
			(0.038)	
Student Course Grade/Term GPA			0.950***	
			(0.158)	
10–20 years teaching experience			−0.510***	
			(0.070)	
20+ years teaching experience			−0.160***	
			(0.040)	
Observations	2,681	2,653	2,376	2,653
R-squared	0.019	0.026	0.077	0.028
Discipline Fixed Effects	YES	YES	YES	YES

Dependent variable is the average positivity rating from three external evaluators. Robust standard errors in parentheses. ***$p < .01$, **$p < .05$, *$p < .1$.

Looking within study group 1, Figure 1 shows that on average female faculty receive less positive feedback than male faculty, and being a harsh grader or having large classes is also correlated with less positive teaching evaluations. The overall gender difference in positivity between male and female faculty is 0.19 points on a 5-point scale, but these gender gaps are exacerbated for being a harsh grader (defined as grading two standard deviations below average), with a gender difference of 0.31, or for those with large classes (classes that are two standard deviations above the average), with a gender difference of 0.36 points. In interpreting the magnitude of these effects, it is important to note that they are calculated holding all other characteristics at the average. If a faculty member has multiple characteristics associated with less positivity, the effects are additive. Similar patterns emerge for study groups 2 and 3, with consistent gender gaps and penalties for being a harsh grader or teaching a larger class.

When comparing across study groups, Table 4 shows that students in group 3, who completed evaluations at the beginning of the following semester are more likely to give positive evaluations relative to students in both groups 1 and 2, who completed their evaluations at the end of that semester, though we do not find a significant difference between groups 1 and 2. The more positive feedback in group 3 is composed of two different effects. First, students who did poorly in the course (as reflected in lower grades) are less likely to complete evaluations when asked for feedback at the beginning of the next semester. Second, when we focus only on students who did well in the course (receiving an A- or better), those students are also more positive, suggesting that sample selection is not the sole reason for this result, but that students who have time to reflect on the course become more positive. However, the increased positivity of

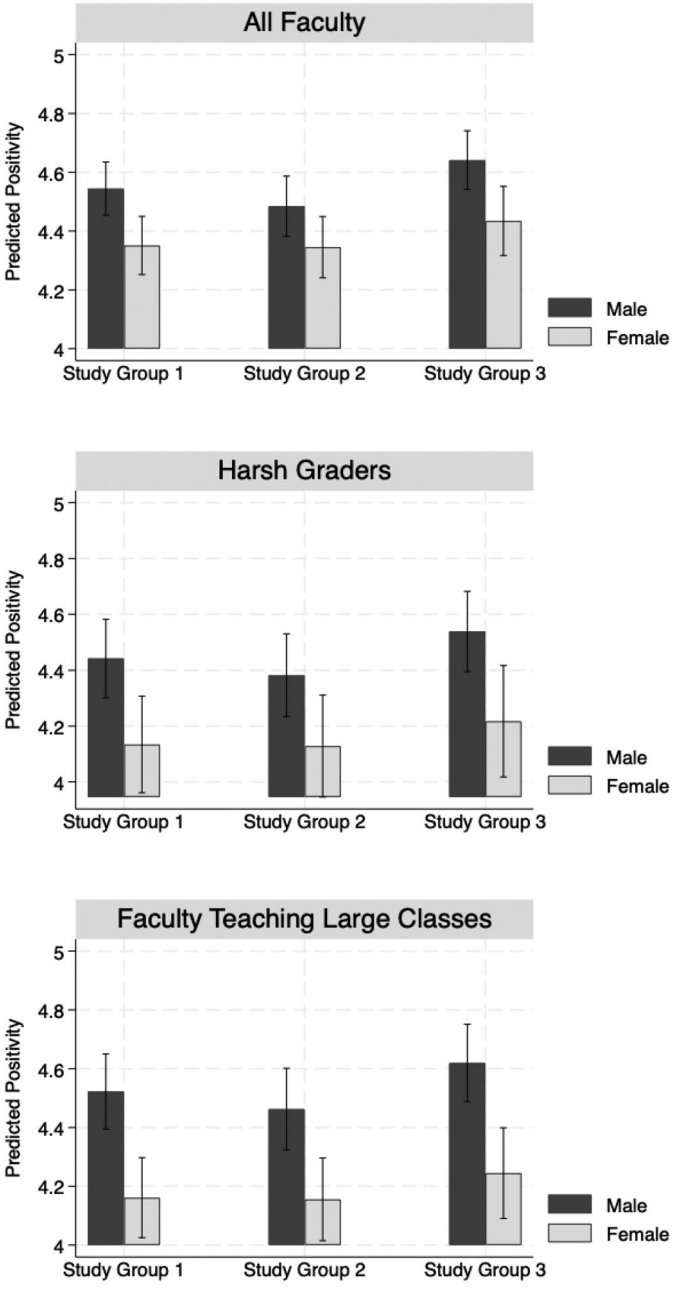

Figure 1. Predicted positivity of student evaluations, by faculty characteristic. Prediction equation is specification reported in Column 4 of Table 4 and includes indicators for study groups and their interaction with the female faculty indicator, average relative grades of faculty and interaction with female faculty indicator, class size and interaction with female faculty indicator, and broad discipline fixed effects. Standard error bands are plus/minus two standard errors. Low grading faculty have average relative grades (average of student course grade/student term GPA for all students) that are two standard deviations below the average. Large classes are two standard deviations above the average. All other characteristics are entered into the prediction equation at their averages.

comments provided by students that complete evaluations at a later time does not reduce the penalties female faculty face. In a context in which faculty are compared to each other using student feedback, the more positive evaluations that resulted from the delayed timing would not eliminate gender bias.

Table 5. Regression results for specificity.

	(1)	(2)	(3)	(4)
Study Group 2	−0.042	−0.041	−0.088*	−0.044
	(0.072)	(0.072)	(0.050)	(0.071)
Study Group 3	−0.114	−0.113	−0.191***	−0.120
	(0.080)	(0.080)	(0.058)	(0.080)
Class Size	−0.001	−0.000	−0.004**	0.005*
	(0.002)	(0.002)	(0.002)	(0.003)
Female Faculty	0.332***	2.442***	0.125**	2.728***
	(0.073)	(0.742)	(0.049)	(0.755)
Study Group 2*Female Faculty	−0.118	−0.137		−0.135
	(0.098)	(0.099)		(0.098)
Study Group 3*Female Faculty	−0.141	−0.142		−0.134
	(0.111)	(0.111)		(0.111)
Avg (Grade/Term GPA) all students		0.735	−1.360***	0.817
		(0.592)	(0.483)	(0.594)
Female Faculty*Avg(Grade/Term GPA) all students		−2.176***		−2.255***
		(0.769)		(0.770)
Female Faculty*Class Size				−0.009**
				(0.004)
Pell Grant Recipient			0.028	
			(0.071)	
White Student			0.251***	
			(0.064)	
Female Student			0.464***	
			(0.057)	
Student Course Grade/Term GPA			0.014	
			(0.172)	
10–20 years teaching experience			0.130*	
			(0.077)	
20+ years teaching experience			−0.128**	
			(0.051)	
Observations	2,681	2,653	2,376	2,653
R-squared	0.022	0.025	0.075	0.027
Discipline Fixed Effects	YES	YES	YES	YES

Dependent variable is the average specificity rating from three external evaluators. Robust standard errors in parentheses. ***$p < .01$, **$p < .05$, *$p < .1$.

In additional results shown in Table 4, we also find that older, more experienced faculty receive less positive evaluations than younger, more recently hired individuals; within a given faculty member's classes, female students provide less positive feedback than male students; and students who earn higher grades relative to their term GPAs provide more positive feedback (Table 4 Column 3). Our findings about the effect of these faculty and student characteristics on the content of qualitative comments largely echo those of studies focused on numeric scores (McPherson 2006; Smith et al. 2007; Spooren 2010). However, in results not shown here, we find that changes to the method and timing at which student feedback is solicited does not reduce the impact of these factors: none of these effects varied significantly across study groups.

Next, we turn to the analysis of the specificity of student evaluations. Regressions using specificity as the dependent variable are shown in Table 5. Again, Columns 1–4 show specifications with different interactions and control variables included. Figure 2 illustrates the results from the specification of Column 4 of Table 5, which includes interactions between faculty gender and study group, grading, and class size. Given the negative relationship between positivity and specificity, we would expect the gender difference in specificity to go in the opposite direction of the difference in positivity, and the figure indeed confirms this. Within study group 1, female faculty receive feedback that is significantly *more* specific than male faculty, with Figure 2 showing predicted specificity levels of 3.3 and 3.0 (on a scale of 1–5), respectively. There is also an interaction between gender and faculty grading patterns, with the gender disparity in specificity being particularly pronounced for faculty who give out low grades. Specifically, we see that harsh grading increases specificity for female faculty (mean of 3.5) but decreases specificity for male faculty

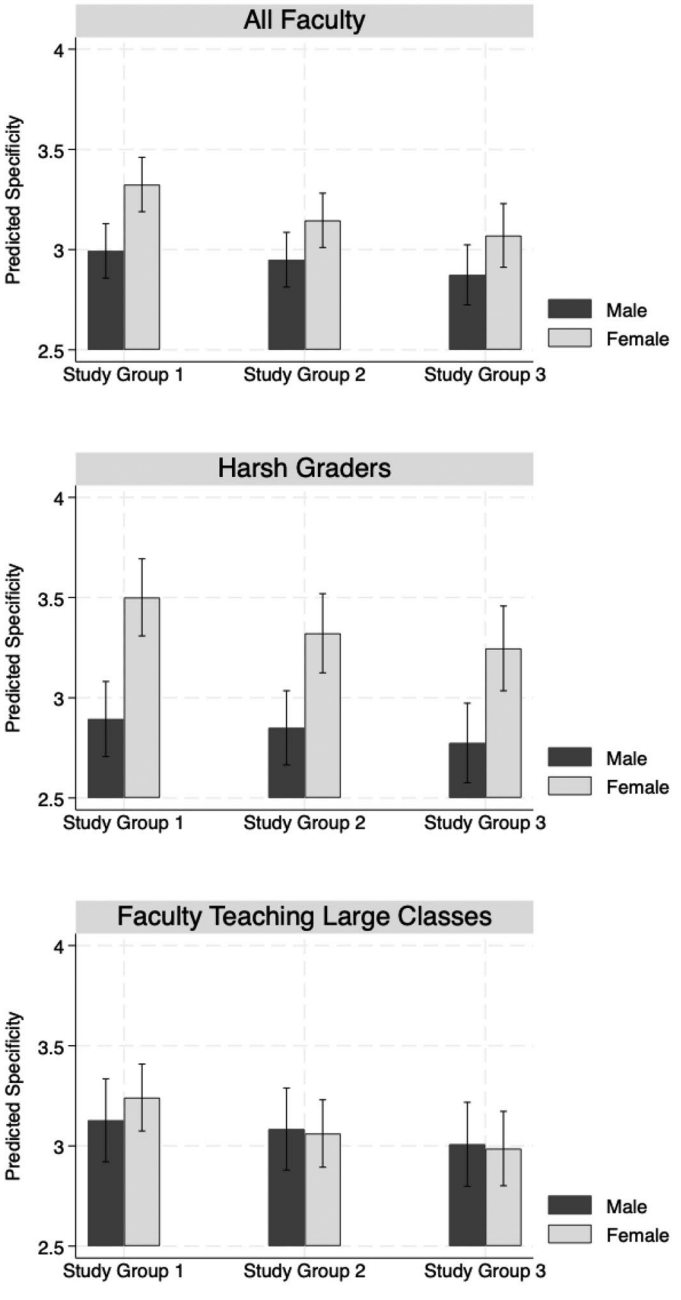

Figure 2. Predicted specificity of student evaluations, by faculty characteristic. Prediction equation is specification reported in Column 4 of Table 5 and includes indicators for study groups and their interaction with the female faculty indicator, average relative grades of faculty and interaction with female faculty indicator, class size and interaction with female faculty indicator, and broad discipline fixed effects. Standard error bands are plus/minus two standard errors. Low grading faculty have average relative grades (average of student course grade/student term GPA for all students) that are two standard deviations below the average. Large classes are two standard deviations above the average. All other characteristics are entered into the prediction equation at their averages.

(mean of 2.9). Meanwhile, there are negligible gender disparities for those teaching large classes. Similar patterns exist within study groups 2 and 3, with female faculty generally receiving more specific feedback, and particularly so when they have low average grades. Looking across study

Table 6. Marginal effect of being female.

Dependent variable	Baseline model, with faculty gender × study group interaction (Column 1)	Including faculty gender × grading interaction (Column 2)	Including additional student and faculty characteristics (Column 3)	Including faculty gender × study group, grading, and class size interactions (Column 4)
Positivity	−0.189***	−0.193***	−0.183***	−0.193***
Specificity	0.332***	0.331***	0.125**	0.325***

***$p < .01$, **$p < .05$. Column headings refer to specifications in Tables 4 and 5. Evaluated for study group 1. Column 1 includes study group indicators and interaction with female faculty indicator, class size, and broad discipline fixed effects. Column 2 adds an interaction with female faculty indicator and grading standards, column 3 adds student characteristics and instructor experience, column 4 removes student characteristics and instructor experience, but adds interaction with class size and female faculty indicator.

groups, Table 5 also shows that students in study group 1 provide more specific feedback than those in the other two groups who complete the two-question evaluation.

In Table 6, we summarize the marginal effects of being female on both positivity and specificity across several different specifications that are reported in Tables 4 and 5. The estimated effects are consistent across specifications, with a slightly lower marginal effect estimated when student characteristics are also included (Column 3 of Tables 4 and 5). Given the negative correlation between specificity and positivity, the results displayed in Figures 1 and 2 and Table 6 provide consistent evidence of a gender gap. Some research on implicit bias suggests that individuals may mask biased decision making by arbitrarily inflating the importance of the characteristics of a specific individual to justify a biased evaluation (Norton, Vandello, and Darley 2004; Uhlmann and Cohen 2005). The greater specificity of evaluation of female instructors may be a symptom of that process.

Because constructive feedback is almost always specific, the results for constructiveness mirror the results for specificity, including the finding that female faculty receive feedback that was rated as more constructive by the external evaluators. We also find that White students and female students provide more constructive feedback than other students, all else being equal. Full regression results for predictors of constructiveness can be found in Appendix Table A2.

At face value, more specific and constructive feedback could benefit female faculty if the feedback is used exclusively for formative purposes. However, at most institutions, student feedback is an important part of the summative evaluation process and affects faculty career progression. To the extent that female faculty respond to this feedback to generate evaluations that are as positive as those of male counterparts, it could require female faculty members to either work harder or award higher grades. Neither of these responses is a desirable outcome, with the former having implications for the professional development of female faculty members and the latter for student learning. Taking the results on the interaction between gender and grades together across all outcomes, we see that female faculty receive evaluations that are qualitatively similar to male faculty only when their grades are especially high relative to grades their students are receiving in their other classes.

As a final step in our analysis, we unblinded the study groups for the external evaluators and asked them to comment on the usefulness of the feedback in each study group as a whole, for the purposes of both summative and formative evaluation. The evaluators noted that the different prompts elicited different types of feedback. In particular, in the open-ended, two-question survey, student feedback focused more on students' own priorities and experiences, while the more directed six-question survey focused on specific teaching practices. The evaluators pointed out that the more directed six-question survey has the advantage of highlighting specific adjustments to be made in the classroom for which there is some consensus, and the more open-ended, two-question survey has the advantage of helping instructors understand the diversity of experiences and expectations of students in the course which could give insight into broader changes that would improve the student learning experience. These evaluator comments give context to the results of our statistical analysis in which we find that students in study group 1 provided more specific and constructive feedback.

Discussion

Despite the mounting evidence that shows how current practices in gathering student feedback of teaching are flawed and prone to bias, little research has been done to assess alternative methods for collecting feedback. Existing work has focused primarily on informational treatments that increase student awareness about the potential for implicit bias, with mixed results (e.g. Key and Ardoin 2019a, 2019b; Peterson et al. 2019; Boring and Philippe 2021). Our study builds on this scholarship by combining an informational treatment with changes in the instrument and timing at which feedback is solicited from students—interventions designed to mitigate situational factors that make people prone to reliance on stereotypes. In contrast to existing work, we test interventions to mitigate bias in qualitative comments, which have the potential to provide important insights into students' classroom experiences, but which also exhibit 'the clearest evidence of gender bias' (Kreitzer and Sweet-Cushman 2022, 79).

Our experiment, conducted at a selective liberal arts college in the United States, randomly assigned students in 210 courses taught by 40 faculty members to one of three groups: a control group who completed an evaluation that consisted of six directed questions at the end of the semester; a treatment group with an alternative prompt with more open ended questions, also solicited at the end of the semester; and a second treatment group with the alternative questions, solicited at the beginning of the *following* semester. Expert external readers who were recruited to read and assess the nature of the feedback found that the set of six directed questions were rated as eliciting more specific and more constructive feedback than the more open-ended alternative prompt. The treatment group that completed evaluations at the beginning of the following semester provided more positive feedback, a finding that is likely due both to selection bias as well as students who performed well in the class providing more positive feedback after a time of reflection.

We find that neither intervention significantly reduced the bias against female faculty. We had hypothesized that the use of a prompt that alerted students to the potential for implicit bias and asked them to evaluate faculty in relation to criteria they articulated for effective teaching, as well as delaying the solicitation of evaluations until a less stressful time for students, would reduce bias. However, we did not find evidence for this hypothesis. Female faculty received qualitatively different feedback than male faculty, and the magnitude of the differences was similar across all study groups. Overall, these findings are consistent with prior literature (e.g. Uttl, White, and Gonzalez 2017; Kreitzer and Sweet-Cushman 2022) in finding strong evidence that the characteristics of students, instructors, and courses impact the nature of student feedback. Across all study groups, female faculty, harsher graders, and those teaching larger courses receive less positive student evaluations, while female students and students who perform worse in the course relative to their overall average grades give less positive feedback.

The failure of these interventions to reduce bias contrasts with the findings two recent studies, which found that informing students about gender bias in student evaluations reduced it (e.g. Boring and Philippe 2021; Peterson et al. 2019). There are several potential explanations for the differences: we test a broader intervention, which pairs an informational treatment with changes to the questions asked and timing at which feedback is solicited; we focus on bias in qualitative evaluation questions, rather than numeric ratings; and we conduct our experiment in a different institutional context, with students across a more diverse set of courses, than existing studies.

As we summarize our conclusions, we also recognize some limitations in our study. Because our intervention for groups 2 and 3 added a preamble about potential for implicit bias and also changed the questions given to students, it is not possible to disentangle the effects of each of these separately. The overall null effect of these interventions on gender bias could be due to insignificant effects of each component or opposing effects of equal magnitude. We also acknowledge the potential for Hawthorne effects, where students would be reluctant to participate and/or behave differently because they have been notified that they are participating in a study. However, when given the choice to opt out of the study, less than 1% of the students did so. Moreover,

Hawthorne effects would only be problematic if they disproportionately affected students in one group relative to another, and there is not necessarily a reason to believe that would be the case.

Overall, our findings highlight the difficulty of removing bias from student feedback. Future research should continue to test the effectiveness of alternative modes of gathering student feedback. Both faculty and administrators would benefit from a better understanding how and to what extent gender bias in evaluations can be mitigated.

Acknowledgments

The research reported in this article was made possible, in part, by a grant from the Spencer Foundation (#202300226). The views expressed are those of the authors and do not necessarily reflect the views of the Spencer Foundation. We are grateful to Melissa Eblen-Zayas, Manisha Goel, Suzanne Keen, Ngonidzashe Munemo, Adriana Signorini, and Michael Sprague for their critical help in implementing the experiment. Aaron Strong and Keelah Williams provided helpful comments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Ann L. Owen http://orcid.org/0000-0001-6513-2491
Erica De Bruin http://orcid.org/0000-0003-4168-9898
Stephen Wu http://orcid.org/0000-0003-4640-0221

Data availability statement

Anonymized data and Stata code to reproduce tables are available at https://doi.org/10.7910/DVN/SFPH9G. The posted data has been transformed to protect educators' confidentiality. Class size data is binned in quartiles to protect the identity of educators who could be identified with more precise information.

References

Alhija, F. N. A., and B. Fresko. 2009. "Student Evaluation of Instruction: What Can Be Learned from Students' Written Comments?" *Studies in Educational Evaluation* 35 (1): 37–44. doi:10.1016/j.stueduc.2009.01.002.

Appleton, L. 2018. "Qualitative Methods for Engaging Students in Performance Measurement." *Information and Learning Science* 119 (1/2): 64–76. doi:10.1108/ILS-09-2017-0093.

Aragón, O. R., E. S. Pietri, and B. A. Powell. 2023. "Gender Bias in Teaching Evaluations: The Causal Role of Department Gender Composition." *Proceedings of the National Academy of Sciences of the United States of America* 120 (4): e2118466120. doi:10.1073/pnas.2118466120.

Aronson, E. 2008. "Social Cognition." In *Social Animal*, edited by E. Aronson. 10th ed, 117–180. New York, NY: Worth Publishers.

Axt, J. R., G. Casola, and B. A. Nosek. 2019. "Reducing Social Judgment Biases May Require Identifying the Potential Source of Bias." *Personality & Social Psychology Bulletin* 45 (8): 1232–1251. doi:10.1177/0146167218814003.

Bedard, K., and P. Kuhn. 2008. "Where Class Size Really Matters: Class Size and Student Ratings of Instructor Effectiveness." *Economics of Education Review* 27 (3): 253–265. doi:10.1016/j.econedurev.2006.08.007.

Berezvai, Z., G. D. Lukáts, and R. Molontay. 2021. "Can Professors Buy Better Evaluation With Lenient Grading? The Effect of Grade Inflation on Student Evaluation of Teaching." *Assessment & Evaluation in Higher Education* 46 (5): 793–808. doi:10.1080/02602938.2020.1821866.

Boring, A. 2017. "Gender Biases in Student Evaluations of Teaching." *Journal of Public Economics* 145: 27–41. doi:10.1016/j.jpubeco.2016.11.006.

Boring, A., and A. Philippe. 2021. "Reducing Discrimination in the Field: Evidence From an Awareness Raising Intervention Targeting Gender Biases in Student Evaluations of Teaching." *Journal of Public Economics* 193: 104323. doi:10.1016/j.jpubeco.2020.104323.

Boring, A., K. Ottoboni, and P. B. Stark. 2016. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." *ScienceOpen Research*. doi:10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

Carrell, S. E., and J. E. West. 2010. "Does Professor Quality Matter? Evidence From Random Assignment of Students to Professors." *Journal of Political Economy* 118 (3): 409–432. doi:10.1086/653808.

Eslami, S. P., M. Ghasemaghaei, and K. Hassanein. 2018. "Which Online Reviews Do Consumers Find Most Helpful? A Multi-Method Investigation." *Decision Support Systems* 113: 32–42. doi:10.1016/j.dss.2018.06.012.

Ghasemaghaei, M., S. P. Eslami, K. Deal, and K. Hassanein. 2018. "Reviews' Length and Sentiment as Correlates of Online Reviews' Ratings." *Internet Research* 28 (3): 544–563. doi:10.1108/IntR-12-2016-0394.

Hoon, A., E. Oliver, K. Szpakowska, and P. Newton. 2015. "Use of the 'Stop, Start, Continue' Method Is Associated With the Production of Constructive Qualitative Feedback by Students in Higher Education." *Assessment & Evaluation in Higher Education* 40 (5): 755–767. doi:10.1080/02602938.2014.956282.

Johnson, T. J., R. W. Hickey, G. E. Switzer, E. Miller, D. G. Winger, M. Nguyen, R. A. Saladino, and L. R. M. Hausmann. 2011. "The Impact of Cognitive Stressors in the Emergency Department on Physician Implicit Racial Bias." *Academic Emergency Medicine* 23 (3): 297–305. doi:10.1111/acem.12901.

Katz, I. 2014. *Stigma: A Social Psychological Analysis*. New York, NY: Psychology Press.

Key, E. M., and P. Ardoin. 2019a. "Students Rate Male Instructors More Highly Than Female Instructors. We Tried to Counteract That Hidden Bias." *The Washington Post*, August 20. https://www.washingtonpost.com/politics/2019/08/20/students-rate-male-instructors-more-highly-than-female-instructors-we-tried-counter-that-hidden-bias/.

Key, E. M., and P. Ardoin. 2019b. "Gender Bias in Teaching Evaluations: What Can Be Done?" Paper Presentation. Southern Political Science Association in Austin, Texas.

Kreitzer, R. J., and J. Sweet-Cushman. 2022. "Evaluating Student Evaluations of Teaching: A Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform." *Journal of Academic Ethics* 20 (1): 73–84. doi:10.1007/s10805-021-09400-w.

Lindahl, M. W., and M. L. Unger. 2010. "Cruelty in Student Teaching Evaluations." *College Teaching* 58 (3): 71–76. doi:10.1080/87567550903253643.

Ma, D. S., J. Correll, B. Wittenbrink, Y. Bar-Anan, N. Sriram, and B. A. Nosek. 2013. "When Fatigue Turns Deadly: The Association Between Fatigue and Racial Bias in the Decision to Shoot." *Basic and Applied Social Psychology* 35 (6): 515–524. doi:10.1080/01973533.2013.840630.

MacNell, L., A. Driscoll, and A. N. Hunt. 2015. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40 (4): 291–303. doi:10.1007/s10755-014-9313-4.

McPherson, M. A. 2006. "Determinants of How Students Evaluate Teachers." *The Journal of Economic Education* 37 (1): 3–20. doi:10.3200/JECE.37.1.3-20.

Mengel, F., J. Sauermann, and U. Zölitz. 2017. *Gender Bias in Teaching Evaluations*. IZA Discussion Papers, No. 11000. Bonn: Institute of Labor Economics (IZA).

Miles, P., and D. House. 2015. "The Tail Wagging the Dog: An Overdue Examination of Student Teaching Evaluations." *International Journal of Higher Education* 4 (2): 116–126. doi:10.5430/ijhe.v4n2p116.

Mitchell, K. M. W., and J. Martin. 2018. "Gender Bias in Student Evaluations." *Political Science & Politics* 51 (03): 648–652. doi:10.1017/S104909651800001X.

Nadler, J. T., M. R. Lowery, J. Grebinoski, and R. G. Jones. 2014. "Aversive Discrimination in Employment Interviews: Reducing Effects of Sexual Orientation Bias With Accountability." *Psychology of Sexual Orientation and Gender Diversity* 1 (4): 480–488. doi:10.1017/S104909651800001X.

Norton, M., J. Vandello, and J. Darley. 2004. "Casuistry and Social Category Bias." *Journal of Personality and Social Psychology* 87 (6): 817–831. doi:10.1037/0022-3514.87.6.817.

Peterson, D. A. M., L. A. Biederman, D. Andersen, T. M. Ditonto, and K. Roe. 2019. "Mitigating Gender Bias in Student Evaluations of Teaching." *PLOS One* 14 (5): e0216241. doi:10.1371/journal.pone.0216241.

Reja, U., K. L. Manfreda, V. Hlebec, and V. Vehovar. 2003. "Open-Ended vs. Close-Ended Questions in Web Questionnaires: Developments in Applied Statistics." *Developments in Applied Statistics* 19 (1): 159–177.

Rivera, L. A., and A. Tilcsik. 2019. "Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation." *American Sociological Review* 84 (2): 248–274. doi:10.1177/0003122419833601.

Smith, S. W., J. H. Yoo, A. C. Farr, C. T. Salmon, and V. D. Miller. 2007. "The Influence of Student Sex and Instructor Sex on Student Ratings of Instructors: Results from a College of Communication." *Women's Studies in Communication* 30 (1): 64–77. doi:10.1080/07491409.2007.10162505.

Snyder, M. L., R. E. Kleck, A. Strenta, and S. J. Mentzer. 1979. "Avoidance of the Handicapped: An Attributional Ambiguity Analysis." *Journal of Personality and Social Psychology* 37 (12): 2297–2306. doi:10.1037/0022-3514.37.12.2297.

Spooren, P. 2010. "On the Credibility of the Judge: A Cross-Classified Multilevel Analysis on Student Evaluations of Teaching." *Studies in Educational Evaluation* 36 (4): 121–131. doi:10.1016/j.stueduc.2011.02.001.

Steyn, C., C. Davies, and A. Sambo. 2018. "Eliciting Student Feedback for Course Development: The Application of a Qualitative Course Evaluation Tool Among Business Research Students." *Assessment & Evaluation in Higher Education* 44 (1): 11–24. doi:10.1080/02602938.2018.1466266.

Stroebe, W. 2020. "Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis." *Basic and Applied Social Psychology* 42 (4): 276–294. doi:10.1080/01973533.2020.1756817.

Uhlmann, E. L., and G. Cohen. 2005. "Constructed Criteria: Redefining Merit to Justify Discrimination." *Psychological Science* 16 (6): 474–480. doi:10.1111/j.0956-7976.2005.01559.x.

Uttl, B., C. A. White, and D. W. Gonzalez. 2017. "Meta-Analysis of Faculty's Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related." *Studies in Educational Evaluation* 54: 22–42. doi:10.1016/j.stueduc.2016.08.007.

Wallace, S. L., A. K. Lewis, and M. D. Allen. 2019. "The State of the Literature on Student Evaluations of Teaching and an Exploratory Analysis of Written Comments: Who Benefits Most?" *College Teaching* 67 (1): 1–14. doi:10.1080/87567555.2018.1483317.

Appendix

Table A1. Balance tests.

	Standardized difference of covariate means		
	Group 1 vs. 2	Group 1 vs. 3	Group 2 vs. 3
Pell Grant Recipient	−0.0455	−0.0662	−0.0052
White Student	−0.0023	−0.0497	0.0276
Female	−0.0434	0.0100	−0.0494
Course Grade/Term GPA	0.0224	−0.0151	0.0318
Female Faculty	0.0059	−0.0228	0.0197
5–10 years experience	−0.0017	0.0119	−0.0088
10–20 years experience	−0.0235	0.0142	−0.0320
20+ years experience	0.0057	−0.0006	0.0061
Avg (Grade/Term GPA) all students	−0.0083	−0.0243	0.0064
Class Size	−0.0341	0.0870	−0.0871

Table A2. Regression results for constructiveness.

	(1)	(2)	(3)	(4)
Study Group 2	−0.113**	−0.113**	−0.131***	−0.112**
	(0.0482)	(0.0483)	(0.0352)	(0.0483)
Study Group 3	−0.180***	−0.179***	−0.213***	−0.177***
	(0.0508)	(0.0507)	(0.0387)	(0.0508)
Class Size	−0.00141	−0.00129	−0.00288**	−0.00285
	(0.00118)	(0.00118)	(0.00122)	(0.00205)
Female Faculty	0.197***	1.745***	0.112***	1.669***
	(0.0489)	(0.485)	(0.0337)	(0.490)
Study Group 2*Female Faculty	−0.0623	−0.0698		−0.0703
	(0.0665)	(0.0670)		(0.0671)
Study Group 3*Female Faculty	−0.0706	−0.0713		−0.0736
	(0.0719)	(0.0724)		(0.0725)
Avg (Grade/Term GPA) all students		0.449	−0.654**	0.427
		(0.393)	(0.333)	(0.393)
Female Faculty*Avg(Grade/Term GPA) all students		−1.600***		−1.579***
		(0.499)		(0.500)
Female Faculty*Class Size				0.00233
				(0.00240)
Pell Grant Recipient			0.0108	
			(0.0446)	
White Student			0.161***	
			(0.0384)	
Female Student			0.249***	
			(0.0348)	
Student Course Grade/Term GPA			−0.202*	
			(0.120)	
10–20 years teaching experience			0.205***	
			(0.0573)	
20+ years teaching experience			0.00432	
			(0.0371)	
Observations	2681	2653	2376	2653
R-squared	0.029	0.033	0.069	0.033
Discipline Fixed Effects	YES	YES	YES	YES

Dependent variable is the average constructiveness rating from three external evaluators. Robust standard errors in parentheses. ***$p < .01$, **$p < .05$, *$p < .1$.