A Lightweight Intervention to Decrease Gender Bias in Student Evaluations of Teaching

Susan Fisk
Department of Sociology
Kent State University
Kent, Ohio, USA
sfisk@kent.edu

Kathryn T. Stolee
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
ktstolee@ncsu.edu

Lina Battestilli
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
lbattestilli@ncsu.edu

***Abstract*—Women are underrepresented as instructors in engineering, computing, and technology classes. One factor that disadvantages women in the classroom are student evaluations of teaching (SETs), as research finds they contain significant gender bias. This may contribute to the dearth of women in computing education, as SETs are used in decisions about contract renewals, hiring, tenure, and promotion. The double-bind is one cause of gender bias in SETs, meaning that it is more difficult for women than for men in leadership positions (such as being a professor) to be perceived as both competent and likable. We examine a lightweight intervention's impact on gender bias caused by the double-bind. Specifically, we conducted a field experiment in which the woman professor of a CS1 class for non-majors gave students in the intervention condition additional, positive exam feedback via email. We hypothesized this would increase students' perceptions of the professor's likability, which would then increase her SETs. We find that the intervention increased top-performing students' ratings of the professors' likability. We also find that the professor received significantly higher SETs the semester she sent the intervention emails. While women should not have to alter their behavior to accommodate students' gender biases, this intervention may be a useful survival strategy for women impacted by gender bias in SETs.**

Keywords— gender bias in teaching, CS1 education, student evaluations of teaching

I. Introduction

Despite efforts to increase the number of women in STEM (Science, Technology, Engineering, and Math), women constitute only 20.8% of Computer Science (CS) faculty across all faculty positions [1]. While many factors contribute to this dearth of women, student evaluations of teaching (SETs) are one source of disadvantage, as they contain significant gender bias [2] [3]. Gender bias in SETs may contribute to women's under-representation in engineering, computing, and technology classrooms, as SETs are used in in decisions about contract renewals, hiring, tenure, and promotion [4].

In an effort to decrease gender bias in SETs, we evaluate the effect of a lightweight intervention. Students in the intervention group received their exam score in an email from the professor with additional, positive feedback that varied based on their exam performance. Top-performers (those with an exam score in the top 50%) in the intervention condition were explicitly told that they had an above-average exam performance and were doing a good job. Bottom-performers (those with an exam score in the bottom 50%) were given positive feedback about their ability to improve and information on resources to help them do so. Students in the control condition received an email with just their score (with no additional feedback or information). We hypothesized that this positive feedback would cause students in the intervention condition to view the woman professor as more likable, and that her SETs would be improved by these increased perceptions of likeability. This is because research finds a strong positive correlation between likability and SETs [5]. Given that women in leadership positions (such as professors) often face a double-bind in which observers fault them for seeming either inadequately nice or inadequately competent [6], this intervention could help decrease likability bias against women professors.

As a disclaimer, the long-term efficacy of this intervention is limited because it does not decrease systematic gender bias in SETs. We also strongly advise against the mandated use of this intervention, as it places an additional burden on women. Despite its limitations and potential for misuse, we report this intervention because it is easier to implement than other survival strategies used by women to combat gender bias (for instance, over-preparation [7]). As such, this intervention may be a useful survival strategy for women at the mercy of SETs (e.g., assistant professors, adjuncts) within institutional settings that are unwilling to make systematic changes to combat gender bias in SETs.

II. Related Work

A. SETs and Bias Against Women Instructors

In higher education, SETs are frequently used in hiring and personnel decisions [4]. However, a growing body of research finds that they are biased against women [4]. For instance, experimental work in online teaching settings has found that students rate instructors they believe to be men higher than instructors they believe to be women, regardless of the instructor's actual gender [3]. Moreover, a natural experiment found that women receive lower SETs by large and statistically significant amounts, even controlling for learning [2]. These effects vary by student gender, with students who are men

978-1-7281-7172-2/20$31.00 ©2020 IEEE

tending to give higher SETs to instructors who are men than to instructors who are women [8]. Thus, in male-dominated fields like computing (in which a majority of students are men), women instructors are likely to be particularly disadvantaged by SETs.

B. The Double-Bind and Gender Bias in SETs

One contributing factor to gender bias in SETs is the double-bind, a dilemma often faced by women leaders in which they can be perceived as either likable but not competent, or competent but not likable. Gender stereotypes drive this effect, as commonly-held beliefs about gender assert that women should be warm, selfless, and nice, while men should be assertive, bold, and agentic [6]. Thus, the gender stereotypes about how men should act line up neatly with societal expectations for leaders, while the gendered expectations for women are in tension with how society believes that leaders should behave [6]. So when women leaders behave in accordance with societal expectations of leaders, they are seen as insufficiently nice. But when they behave in accordance with the gendered expectations held for women, they are seen as inadequately competent leaders. The double-bind is challenging for women academics because the role of instructor often requires giving negative feedback to students. And indeed, students rate difficult graders more poorly when they are women [9].

C. Likability Interventions and the Double-Bind

Some research has found that women leaders can overcome the double-bind if they act in a competent manner while demonstrating traits consistent with the gender stereotypical expectations of women (e.g., nice, communal, and group-orientated) [7]. For instance, backlash against women who negotiate is negated when women negotiate for others [10].

Thus, we suspect that women instructors who engage in warm, friendly behavior towards their students may be able to overcome the likability bias of the double-bind. We hypothesize that this will improve the SETs of women instructors, as SETs are highly correlated with likability [5] and friendliness towards students has been shown to increase SETs for women instructors but not men instructors [11].

III. Research Questions

We evaluate two research questions:

- **RQ1:** To what degree does additional, positive feedback from the professor delivered via email increase students' perceptions of the woman professor's likability?
- **RQ2:** To what degree does additional, positive feedback from the professor delivered via email increase SETs for a woman professor?

IV. Methods

This study uses two methods to evaluate our research questions. For RQ1, we use data from a controlled A/B study in which half the students got the intervention and half were the control. For RQ2, we use the official University SETs for the semester in which the intervention occurred (considering all students, even those in the control), and compare against a control semester that did not use the intervention at all.[1]

A. Context

This study was conducted in the Fall semester of 2018 in a CS1 course for engineering students (non-majors) at a large public University in the United States. Students took surveys both before (Pretest survey) and after (Posttest survey) their first exam. All students in the Fall 2018 offering of the course were required to complete the Pretest survey at the start of the course.[2] Students were then offered 2 percentage points of extra credit for completing a Posttest survey, which was given after the exam.[3]

B. Assignment of Participants to Intervention Group

After the first exam, students were stratified by exam performance (top 50% or bottom 50%). They were then randomly assigned to either the control or intervention group. While not every student consented to the use of their data for this research, every student was assigned to the control or the treatment group.[4]

1) Control Group Emails: After the exam, students in the control group received an email in which they were only given their numeric grade on the exam followed by information on how to access the survey.

2) Intervention Group Emails: After the exam, students in the intervention group received an email from the professor giving them their numeric grade on the exam, information on how to access the survey, as well as additional feedback that varied based on their exam performance. Top-performers (top 50% of exam scores) in the intervention condition were explicitly told that they had an above-average exam performance and were doing a good job. Bottom-performers (bottom 50% of exam scores) were given positive messaging about their ability to improve and information on resources to help them do so.

C. Metrics

Two metrics were used for the evaluation of the research questions: 1) professor likability (from Pretest and Posttest surveys) and 2) official SETs (administered by the University at the end of every course).

1) Likability of Professor: On both surveys, students were asked, *"How much do you like the instructor of this class?"* and could respond on a 7-point scale (in which 1 = "Greatly dislike," 4 = "Neither like nor dislike," and 7 = "Greatly like"). The mean likability score (across both the Pretest and Posttest survey) was 5.51 with a standard deviation of 1.09. We use linear mixed models to assess the impact of the intervention on student perceptions of professor likability.

[1]Due to the space limitations, full details on the study methods, analyses, and results are available in a technical report [12].

[2]However, students were not required to consent to the use of their data.

[3]All students, independent of test performance and consent for data use, could earn extra credit by completing the surveys.

[4]We did this so that the professor could not know which students had consented to data use.

TABLE I
LINEAR MIXED MODELS WITH REPEATED MEASURES PREDICTING TOP PERFORMING STUDENT RATINGS OF PROFESSOR LIKABILITY

	Coefficient	Standard Error	p-value
Time	0.03	0.13	0.83
Intervention	0.33	0.16	0.04
Intercept	5.66	0.19	0.00

n=148 observations nested in 74 participants.
NOTE: Each model has a random intercept and an AR(1) specification for serial correlation.

2) Student evaluations of teaching: The professor's official SETs were used to assess the impact of the intervention on teaching evaluations. We compared the professor's SETs from Fall 2018 (the semester the intervention occurred) to her Spring 2019 SETs (a semester in which she sent no emails about exam grades). This semester was used because it was most directly comparable to the intervention semester, given its close temporal proximity and the minimal course changes that occurred between the two semesters. For each question in the SETs, students responded on a 5-point scale in which 1 = "strongly disagree" and 5 = "strongly agree". Blank or "not applicable" responses were removed from this analysis.

D. Participants and Response Rates

While there were 185 students in the Fall 2018 class, control and treatment groups were assigned based on the 167 students who consented. However, there was an unequal distribution between the control (67 students) and intervention groups (72 students) because only 139 students completed both the Pretest and Posttest surveys. Thus, we report a response rate of 139/185 (75.1%). Of these 139 students, 74 were classified as top performers (35 control, 39 intervention) and 65 were classified as bottom performers (32 control, 33 intervention). Among students who participated, all identified as either women (29 students) or men (110 students).[5]

For SETs, the response rate was 80/185 (43.2%) for the intervention semester of Fall 2018 and 103/264 (39.0%) for the control semester of Spring 2019.

V. RESULTS

A. RQ1: Impact of Intervention on Professor Likability

Using the data from the Pretest and Posttest surveys from Fall 2018, we find direct evidence that the intervention causes top-performing students to like the professor more. Table I shows the results of the analysis with linear mixed models. Time takes on a value of '1' for the Pretest survey and '2' for the Posttest survey. Intervention takes on a value of '0' for all observations at time 1 (as no students had received the intervention at this time), and takes on a value of '1' at time 2 if the student was in the intervention group. We conduct separate analyses for top and bottom performers, given the differences in the feedback received by these groups.

[5]Gender was balanced across the control and treatment groups. We do not break down the analysis by student gender because student gender did not impact the effect of the intervention.

We find evidence that the intervention increases top-performing student ratings of professor likability by .33 points ($p < .05$). This represents an increase of 5.8% percent, given the average rating of professor likability in the control group was 5.66. While this is a modest increase, it is statistically significant.

We do not find evidence that the intervention increases bottom-performing student ratings of professor likability, with $p = 0.80$, see [12]. Similarly, when considering top performers and bottom performers in aggregate, there is no significant overall effect of the intervention.

B. RQ2: Impact of Intervention on SETs

To determine if the intervention influences SETs, we compare the professor's Fall 2018 SETs (the intervention semester) with her Spring 2019 SETs (the comparison semester). We use a paired t-test in which we treat each of the twelve SET questions as a unit and then use each semester's average value for the question as a repeated measure of the unit. This means that the average SET score received by the professor in the Fall of 2018 (the intervention semester) was 4.13 (with a standard deviation of 0.28 and 12 observations - one for each of the questions), and the average score the professor received in the Spring of 2019 was 3.92 (with a standard deviation of 0.22 and 12 observations - one for each of the questions). This difference was found to be statistically significant, with a t-statistic of -5.84 and a p-value of ¡ 0.001.

VI. DISCUSSION

Women constitute a minority of professors in engineering, computing, and technology courses, and face challenges that their counterparts who are men do not. One of these challenges is that SETs have been found to be biased against women, so much so that the American Sociological Association (ASA) released a statement cautioning against the use of SETs in tenure and promotion cases [4].

In this work, we present a lightweight intervention that appears to decrease gender bias in SETs by mitigating the effects of likability bias against women professors. Although we only found evidence that the intervention increased the top-performing students' perceptions of the professor's likability (RQ1), the positive messaging in the intervention appears to be so effective that the intervention led to significantly higher SETs at the end of the semester (RQ2). Although it might seem unlikely that a single email could have a large impact on SETs, it is well established that a single action can greatly impact observers' attributions and understandings of a person, especially when that action occurs early in the relationship between the person and the observer [13]. Future research should more directly assess the precise mechanisms that caused the intervention email to increase perceptions of the professor's likability and her SETs.

While women should not have to change their behavior to accommodate the gender bias of students, an unfortunate reality is that most institutions of higher learning use SETs to evaluate faculty. This intervention may be helpful to women

who are struggling with the effects of gender bias caused by the double-bind, as this intervention may be easier for them to implement than other behavioral adjustments they already use to circumvent gender bias (for instance, over-preparing for class or carefully curating ones appearance [7]).

Given the important role women professors play in the retention of top students who are women, this intervention may also have important downstream effects on women students in STEM. Carrell finds that while professor gender has little impact on students who are men, it does have a powerful effect on the performance and retention of students in STEM courses who are women, especially top performing women [14]. If this intervention helps retain more women instructors in STEM fields, it may also have the additional benefit of increasing the retention of students in STEM who are women.

A. Other Factors At Play

Were the SETs better during the intervention semester because the class size was smaller? One might argue that SETs were better duing the intervention semester because there were 79 fewer students enrolled in the intervention semester than the control semester. To assess this hypothesis, we compared the SETs from the Spring 2018 semester (a semester in which an identical email intervention occurred, but that did not include the likeability question on the surveys) and the Fall 2017 semester (its closest control semester). In this case, the Spring 2018 enrollment (the intervention semester) was higher than the Fall 2017 enrollment (the control semester). We again find that the professor's SETs were significantly higher ($p = 0.026$) during the intervention semester (Spring 2018) than the control semester (Fall 2017).

Was the quality of instruction higher during the intervention semester? One might argue that the professor was particularly invested in teaching during the semester of the intervention. However, the intervention did not appear to have an impact on the questions from the SET related to professor explanations, enthusiasm, and preparedness, or course materials [12], which leads us to believe the delivery of material, and the materials themselves, were similar between semesters. Moreover, the professor stated that she did not change the course materials between semesters.

Is the improvement in SETs due to higher response rates? There was a higher response rate for SETs in the of Fall 2018 compared to the Spring of 2019 (See Section IV-D), but using a 2-sample test for equality of proportions with a continuity correction, we find that the difference in response rates is not significant ($p = 0.184$).

B. Threats to Validity

1) External Validity: We studied students in a CS1 course for non-majors at a large research University in the United States and results may not generalize to other populations, such as smaller institutions or courses for majors. Results are reported for one professor who is a woman and may not generalize to other women or professors of other genders.

2) Conclusion Validity: The response rates for the SETs were 43.2% and 39.0% for Fall 2018 and Spring 2019, respectively. It is possible that the results may not hold with a higher response rate.

VII. Conclusion

Research finds that SETs contain significant gender bias, and that professors who are women often receive lower evaluations than men for a similar quality of instruction. We examined the effects of a lightweight email intervention that provides positive exam feedback to students. We find evidence that the intervention improves short-term student perceptions of the women professor's likability. We also find evidence that the intervention increases official SETs, providing further evidence of the intervention's effectiveness. While this intervention does not decrease the gender bias of students, this intervention could be a survival strategy used by women to mitigate bias they experience in SETs.

Acknowledgements

We thank the anonymous reviewers for their time and suggestions. This work was supported in part by NSF #1749936.

References

[1] S. Zweben and B. Bizot, "The taulbee survey," *Computing Research Association*, 2018. [Online]. Available: https://cra.org/resources/taulbee-survey/

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

[3] L. MacNell, A. Driscoll, and A. N. Hunt, "What's in a name: Exposing gender bias in student ratings of teaching," *Innovative Higher Education*, vol. 40, no. 4, pp. 291–303, 2015.

[4] A. S. Association, "Statement on student evaluations of teaching," 2019.

[5] D. Feistauer and T. Richter, "Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest," *Studies in Educational Evaluation*, vol. 59, pp. 168–178, 2018.

[6] A. H. Eagly, L. L. C. A. H. Eagly, and L. L. Carli, *Through the labyrinth: The truth about how women become leaders*. Harvard Business Press, 2007.

[7] J. C. Williams and R. Dempsey, *What works for women at work: Four patterns working women need to know*. NYU Press, 2018.

[8] J. A. Centra and N. B. Gaubatz, "Is there gender bias in student evaluations of teaching?" *The journal of higher education*, vol. 71, no. 1, pp. 17–33, 2000.

[9] S. A. Basow and N. T. Silberg, "Student evaluations of college professors: Are female and male professors rated differently?" *Journal of educational psychology*, vol. 79, no. 3, p. 308, 1987.

[10] E. T. Amanatullah and M. W. Morris, "Negotiating gender roles: Gender differences in assertive negotiating are mediated by women's fear of backlash and attenuated when negotiating on behalf of others." *Journal of personality and social psychology*, vol. 98, no. 2, p. 256, 2010.

[11] D. Kierstead, P. D'agostino, and H. Dill, "Sex role stereotyping of college professors: Bias in students' ratings of instructors." *Journal of Educational Psychology*, vol. 80, no. 3, p. 342, 1988.

[12] S. Fisk, K. T. Stolee, and L. Battestilli, "A Lightweight Intervention to Decrease Gender Bias in Student Evaluations of Teaching (Full Version)," North Carolina State University, Department of Computer Science, Tech. Rep., 12 2019. [Online]. Available: ftp://ftp.ncsu.edu/pub/unity/lockers/ftp/csc_anon/tech/2019/TR-2019-9.pdf

[13] K. G. Shaver, *An introduction to attribution processes*. Routledge, 2016.

[14] S. E. Carrell, M. E. Page, and J. E. West, "Sex and science: How professor gender perpetuates the gender gap," *The Quarterly Journal of Economics*, vol. 125, no. 3, pp. 1101–1144, 2010.