A Lightweight Intervention to Decrease Gender Bias in Student Evaluations of Teaching (Full Version)

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

Kathryn T. Stolee, Lina Battestilli
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
{ktstolee, 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. Research suggests that one contributor to gender bias in SETs is the double-bind, 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 ($p < 0.05$). We also find that the professor received significantly higher SETs ($p < 0.001$) the semester she sent the intervention emails, even though the intervention was administered to half the students in the class. While this intervention does not eliminate the gender disparities faced by woman instructors, and while women should not have to alter their behavior to accommodate students' gender biases, it is a promising lightweight intervention that could help women harmed by gender bias in SETs.**

I. Introduction

Despite efforts to increase the number of women in STEM, women remain underrepresented as instructors of engineering, computing, and technology courses. Based on the 2018 Taulbee survey published by the Computing Resource Association (CRA), women constituted only 20.8% of Computer Science (CS) faculty across all faculty positions (e.g., tenure track, teaching, research, postdoc) [21]. While many factors contribute to the dearth of women in these fields, student evaluations of teaching (SETs) are one source of disadvantage, as research finds they contain significant gender bias. For instance, women have been found to receive lower SETs even when learning outcomes are taken into account [6] and experimental work in online settings (in which instructor gender is randomly assigned) finds that instructors perceived to be women receive lower evaluations than otherwise identical instructors perceived to be men [13]. Gender bias in SETs may be a contributing factor 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 on: 1) student perceptions of their professor's likability and 2) SETs. 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 (defined as receiving 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 (defined as receiving 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 that likability is highly positively correlated with SETs [11]. 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 [9], this intervention could help decrease likability bias against women professors.

Thus, the goal of this work is to assess the following research questions:

RQ1: Does additional, positive feedback from the professor delivered via email increase students' perceptions of the woman professor's likability?

RQ2: Does additional, positive feedback from the professor delivered via email increase SETs for a woman professor?

In summary, we answer RQ1 and RQ2 in the affirmative. Top-performing students in the intervention group (who received additional, positive exam feedback from the professor) provided significantly higher ratings of the woman professor's likability compared to students in the control group. Furthermore, the woman professor's SETs were higher the semester that she sent out emails with additional feedback compared to the following semester with standard feedback. Thus, we find preliminary evidence that increased student feedback–when it is positive–can increase the SETs of women professors in computing. Although women should not have to alter their behavior to correct for the gender biases of others, an unfortunate reality is that SETs are used by a majority of institutions of higher education, and that many of these institutions are unwilling to take into account the effects of gender bias in SETs or to make systematic changes to decrease the gender biases that plague SETs. Thus, we hope that this intervention may prove to be a valuable, "survival strategy," for women working to be successful in computing education.

To our knowledge, this is the first study that examines the impact of a lightweight, email intervention on the likability of women professors. While we study one professor in an introductory Computer Science (CS1) course for non-majors, the results are encouraging and suggest future research with other professors in other courses may yield similar results. The rest of this paper is organized as follows. Section II reviews related work in research involving SETs and the double-bind. Section III describes the methods we used to deploy an intervention in which half the class got additional, positive feedback from the professor and measures the impact of the intervention through SETs. The results are in Section IV, followed by discussion in Section V and conclusion in Section VI.

II. Related Work

A. Student Evaluations of Teaching and Bias Against Women Instructors

In higher education, assessments of quality of instruction are frequently used in decisions about hiring, tenure, promotion, contract renewal, and merit raises [4]. While quality of instruction can be evaluated in a multitude of ways (for instance, evaluations from the instructors themselves, peers, or experts), many institutions use SETs because they are inexpensive and simple to collect. These SETs typically take the form of close-ended items (i.e., Likert scales) and are usually administered at the end of the semester. However, using SETs to evaluate quality of instruction is problematic because research suggests that SETs are not good measures of student learning outcomes, and are instead highly correlated with anticipated grades [6].

In addition, a growing body of research finds that they are heavily influenced by factors that are out of the control of the instructor, such as the time of day the course is offered, the race of the instructor, and of particular interest to this study, the gender of the instructor [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 [13]. Moreover, in the aforementioned experimental study, the students even rated the instructor believed to be a woman worse on objective items, such as the promptness of returning graded assignments [14]. Other research using a natural experimental setting has found that women receive lower SETs by large and statistically significant amounts, even controlling for learning [6]. These effects vary by discipline and student gender, with students who are men 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 disadvantages 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 either be perceived as 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 [15]. 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 [9]. 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 often seen as inadequately competent.

The double-bind presents a particular challenge for women in academia because the role of instructor often requires giving negative feedback to students, which can result in a perceived lack of niceness (or perhaps even meanness). And indeed, research finds that giving negative feedback has a worse impact on the perceptions of women than of men [18], and that difficult graders who are women are rated more poorly by students than difficult graders who are men [5]. Moreover, this problem may be compounded in computing classes, as average grades in science, technology, engineering and math (STEM) courses tend to be lower than other university courses [2], with some of the lowest grades on campus being given in introductory STEM courses [16].

C. Likability Interventions and the Double-Bind

Some research has found that women leaders can overcome the double-bind when they demonstrate competence while also working to, "...express interest and concern about the lives of those who work for them," [20, pg.92]. In other words, women can behave in accordance with stereotypes of leaders without experiencing a backlash, so long as they demonstrate traits consistent with the gender stereotypical expectations of women (e.g., being nice, communal, and group-orientated). Another compelling example of this comes from research on negotiation, which finds that backlash against women who negotiate is negated when women negotiate for others [3]. In other words, women can act in an agentic, assertive manner when they negotiate, so long as they are acting in accordance with gendered prescriptions that they are communal and other-oriented.

Given that women instructors in computing courses almost always have undisputed competence relative to their students (as they have generally achieved advanced degrees), it seems probable the aspect of the double-bind that is most disadvantaging to women instructors in computing is the likability bias. Thus, we suspect that women instructors who engage in warm, friendly behavior towards their students may be able to offset likability bias against them. This is supported by research finding that friendliness towards students has been shown to increase SETs for women instructors but not men instructors [12], and research showing that women instructors were judged by students as less likable if they did not extensively interact with them [19].

We hypothesize that having a woman instructor give students feedback – especially feedback that could be perceived as negative – in a positive, friendly manner will cause students to see the professor as more likable. Moreover, we hypothesize that this will also improve SETs, as research finds that SETs are highly correlated with likability [11].

III. Methods

This study uses two methods to evaluate the research questions. For RQ1, we use data from a controlled A/B study in which half a class of 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.

A. Context

This study was conducted in a CS1 course for engineering students (non-majors) at a large public University in the United States. There are two days of lecture (50 minutes each) and one lab (2 hours each) per week, and between 185 to 290 students enroll each semester. This course has been taught by the same woman professor since Fall 2012. We report on the effects of an intervention that took place during the Fall semester of 2018.

In work done concurrently [1], we developed an intervention designed to increase students' CS persistence intentions by providing them with additional feedback about their test performance via email. We conducted the first-round of data collection using a field experiment in the Spring of 2018, and found evidence that among top performers (top 50% of exam scores), the intervention increased women's CS persistence intentions (but not men's). It was then suggested to us that the intervention may have improved women's intentions to persist in computer science because it caused them to like the professor more. Thus, we added a question about professor likability to our second-round of data collection (a field experiment we conducted in the Fall of 2018). We found that student perceptions of professor likability did not have a statistically significant impact on computer science persistence intentions.

We became interested in studying the effect of the intervention on perceptions of the professor's likability because existing research on gender stereotypes suggest that the intervention could increase perceived likability. Moreover, we had already added the question about professor likability to the survey. Despite the fact that we conducted a similar field experiment in the Spring of 2018 (during our first-round of data collection), and despite the fact that the professor also had higher than average teaching evaluations during the first-round of data collection (in the Spring of 2018: see Section V-A), we limit our analysis in this paper to the Fall 2018 data, as that was the only semester that we directly asked students about their perceptions of the professor's likeability.

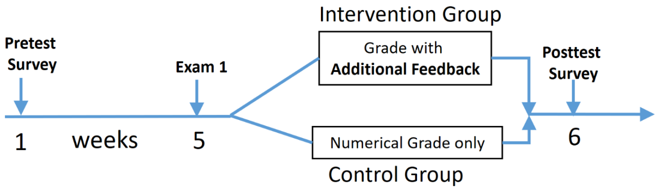

Fig. 1. Timeline of the Experiment

In the concurrent work [1], students took surveys with a number of questions (e.g., about self-assessed CS ability, CS persistence intentions) both before (Pretest survey) and after (Posttest survey) the exam. All students in the Fall 2018 offering of the course were required to complete the Pretest survey at the start of the course.[1] Students were then offered 2 percentage points of extra credit for completing a Posttest survey, which was given after the exam.[2] See Figure 1 for a timeline of the surveys.

B. Assignment of Participants to Intervention Group

After the 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.[3]

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 the survey. The text is shown in Figure 2.

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, 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.

Top performers were also told whether they placed in the top 10%, top 25%, or top 50% of students in the course. At

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

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

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

“You got an XX% on the test.

We are also interested in learning more about your course experiences and career aspirations. If you help us by completing this brief survey by XX date, 2 points of extra credit will be added to your exam score. The survey will only take about 5 minutes to complete, and your responses will help future students in this course. <survey link>”

Fig. 2. Email content for the control group

“You got an {{Grade}}% on {{Test}}! Congratulations! Since average grades in STEM courses tend to be lower than in other university classes, I wanted to make sure that you know that you are a top performer in the class! [You scored in the top 10%, and earned the X highest score in the class!/ Your score places you in the top quarter of all grades on this test!/ You scored better than half of the students in this class!] Keep working hard! I know that you have what it takes to be successful in Computer Science!

We are also interested in learning more about your course experiences and career aspirations. If you help us by completing this brief survey by XX date, 2 points of extra credit will be added to your exam score. The survey will only take about 5 minutes to complete, and your responses will help future students in this course. <survey link>”

Fig. 3. Email content for the intervention group; top 50%

the end of the email message was a GIF of dancing minions,[4] in order to affectively reinforce the positive feedback of the email message, as shown in Figure 3.

Students in the intervention group who scored in the bottom 50% received the email message in Figure 4, which includes positive messaging about their ability to improve and information on resources that may help.

“You got a XX% on the test. Remember that average grades in STEM courses tend to be lower than in other university classes and that many people do not perform well on their first computer science test.

I also believe that if you put in the time and work hard, you can improve the grade on your next test. Research shows that passion, dedication, and self-improvement – and not simply innate talent – are the road to genius and contribution. Indeed, research finds that on average, students who excel at STEM courses spend more time and energy preparing for class, studying, and trying to improve themselves.

Here are some resources that might be helpful to you:

- *Office Hours (led by TAs and <Professor>) and Study Hours (led by Study Hour leaders). Specific times and locations are posted on the Google Calendar on Moodle.*
- *Video of the Lectures (recorded by* $< Professor >$ *for Engineering Online, requires University login)*
- *The online interactive textbook*

Again, I believe that you have what it takes to be successful in this course if you work hard. Please reach out at any time if there is anything else I can do to help you succeed.

We are also interested in learning more about your course experiences and career aspirations. If you help us by completing this brief survey by XX date, 2 points of extra credit will be added to your exam score. The survey will only take about 5 minutes to complete, and your responses will help future students in this course. <survey link>”

Fig. 4. Email content for the intervention group; bottom 50%

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 the Pretest and Posttest 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, as each student gave two ratings of the professors’ likability (one during the Pretest survey before the exam at week 1 time 1, and one during the Posttest survey after the exam at week 6 time 2: see Figure 1). This gives the data a nested structure, in

[4] Minions are cartoon characters from the children’s movie, Despicable Me (as well as its sequels and spin-offs). As defined by Edwards, “Minions are a species of tiny yellow henchmen; they look like unusually dressed Mike and Ike candies. They’re earnestly driven by the desire to serve an evil boss, though they often screw up because they’re selfish, easily distracted, and generally inept. They vary in height, but it’s safe to say they’re between 2 and 3 feet tall,” [10]

TABLE I
PARTICIPANT BREAKDOWN BY INTERVENTION STATUS AND PERFORMANCE

	Top Performers	Bottom Performers	Total
Control	35	32	67
Intervention	39	33	72
Total	74	65	139

which ratings of likability are nested within individual students. Using mixed models allows us to take into account the fact that observations are not independent (as each student contributed two) while still utilizing the full statistical power provided by repeated measures. Given the time-lag between observations, within-group errors were modeled to have an autoregressive structure with a lag of 1.

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 as a comparison 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. Table III lists the twelve questions from the SETs. For each question, 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

For the survey item about likability of the professor, Table I shows the breakdown of participants based on performance and whether they were in the control or intervention group. While there were 185 students in the class, only 167 consented to the use of their data for research. Control and treatment groups were assigned based on the 167 students who consented. However, only 139 students completed both the Pretest and Posttest surveys. This created an unequal distribution between the control and intervention groups (67 vs. 72). We report a response rate of 139/185 (75.1%). Among students who participated, all identified as either women (29 students) or men (110 students).[5]

For student evaluations of teaching, 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.

IV. RESULTS

We address each research question in turn.

A. RQ1: Impact of Intervention on Professor Likability

We first examine the effect of the intervention on perceptions of the professor's likability. Using the data from the Pretest and Posttest surveys, we find direct evidence that the intervention causes top-performing students to like the professor more. Table II 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.

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 $p = 0.80$. Similarly, when considering top performers and bottom performers in aggregate, there is no significant overall effect of the intervention.

RQ1 Summary: The lightweight, positive e-mail intervention increases top-performing students' ratings of professor likability by 0.33 points on a 7-point scale, which is statistically significant ($p < 0.05$).

B. RQ2: Impact of Intervention on Student Evaluations of Teaching

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). On a per-question basis, we compare the responses between semesters using t-tests and report the t-statistic and the p-value. For example, we find significant differences between the semesters for the first question, *The instructor's teaching aligned with the course's learning objectives/outcomes*, with a t-statistic of -3.10 and a p-value of 0.002. Table III presents the full results.

We find that SETs for questions that specifically make reference to the instructor (questions 1–8 in Table III) are higher during the intervention semesters, and the differences are significant for questions 1–2 and 6–8. Specifically, students felt like the feedback they received was more useful (question 6) and that they were treated with more respect (question 7) in the intervention semester compared to the control semester ($p < 0.05$). Additionally, students felt that the professor was more receptive to students outside the classroom (question 2, $p < 0.01$). According to student responses, the quality of instruction in the intervention semester was higher (question 1, $p < 0.01$) and the professor was more effective (question 8, $p < 0.05$).

We did not observe a statistically significant difference in evaluations related to professor enthusiasm (question 4), professor preparation (question 5), or professor explanations (question 3). Further, there were no observed significant differences between semesters for questions related to the

[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.

TABLE II
LINEAR MIXED MODELS WITH REPEATED MEASURES PREDICTING STUDENT RATINGS OF PROFESSOR LIKABILITY

	Top Performers			**Bottom Performers**		
	Coefficient	Standard Error	p-value	Coefficient	Standard Error	p-value
Time	0.03	0.13	0.83	-0.02	0.18	0.93
Intervention	0.33	0.16	0.04	0.06	0.24	0.80
Intercept	5.66	0.19	0.00	5.32	0.27	0.00

Top Performers had n=148 observations nested in 74 participants.
Bottom Performers had n=130 observations nested in 65 participants
NOTE: Each model has a random intercept and an AR(1) specification for serial correlation.

TABLE III
SCORES FOR THE STUDENT EVALUATIONS OF TEACHING; SCORES ON A LIKERT SCALE FROM 1 TO 5 WITH 5 BEING THE HIGHEST.

		Intervention (Fall 2018)			**Control (Spring 2019)**				
	Question	Mean	stdev	n	Mean	stdev	n	t-statistic	p-value
1	The **instructor's** teaching aligned with the course's learning objectives/outcomes	4.40	0.61	80	4.09	0.76	103	-3.10	0.002 **
2	The **instructor** was receptive to students outside the classroom	3.96	0.85	71	3.60	0.98	98	-2.51	0.013 **
3	The **instructor** explained material well.	3.90	0.89	80	3.63	1.08	103	-1.85	0.067 †
4	The **instructor** was enthusiastic about teaching the course	4.29	0.70	80	4.21	0.75	103	-0.69	0.492
5	The **instructor** was prepared for class	4.31	0.70	80	4.20	0.68	103	-1.05	0.294
6	The **instructor** gave useful feedback.	3.94	0.90	78	3.61	1.08	99	-2.22	0.028 *
7	The **instructor** consistently treated students with respect	4.24	0.75	79	3.95	1.00	102	-2.22	0.028 *
8	Overall, the **instructor** was an effective teacher	4.00	0.86	79	3.63	1.08	103	-2.57	0.011 *
9	The course materials were valuable aids to learning	3.90	1.09	80	3.91	1.00	102	0.08	0.940
10	The course assignments were valuable aids to learning	4.38	0.86	80	4.14	0.97	100	-1.71	0.088 †
11	This course improved my knowledge of the subject	4.44	0.71	78	4.36	0.74	102	-0.67	0.504
12	Overall, this course was excellent	3.70	1.00	79	3.87	1.04	102	-1.16	0.248
Average (all questions equally weighted)		4.13	0.28	12	3.92	0.22	12	-5.84	< 0.001 ***

†$p \leq .10$ *$p \leq .05$ **$p \leq .01$ ***$p \leq .001$.

course content (question 9), course assignments (question 10), improvements to student knowledge (question 11), and the excellently of the course (question 12).

To test whether there is an overall difference in SETs between the two semesters, 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. This is reported in the final row of Table III, labeled *Average (all questions weighted equally)*.

It is worth noting that when we center the data using the department averages for each question for each semester, the results of this analysis are the same (if not more favorable for the effect of the intervention). One could argue that mean centering is appropriate to account for overall department trends and/or time effects. However, since the two semesters were adjacent in time (separated by only a winter holiday), in this case we chose the non-centered data as it makes a clearer presentation and reveals the same effect.

RQ2 Summary: Overall student evaluations of teaching were significantly higher in the intervention semester, despite the fact that the intervention was given to only half the students. The biggest differences were seen in questions related to the professor, including her overall effectiveness, the usefulness of her feedback, and treatment of students.

V. 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]. Yet despite their flaws, SETs remain a reality in the lives of most instructors, and are often a metric used in the evaluation of faculty, including in reappointment, tenure, and promotion decisions.

In this work, we present a lightweight intervention that appears to mitigate the effects of likability bias against women professors and decrease gender bias in 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 quality of instruction and do not take into account the effect of gender bias on these evaluations. While this intervention will not fix the double-bind or eradicate gender bias against women, this intervention may be helpful to women who are struggling with the effects of gender bias caused by the double-bind. Indeed, many women already report adjusting their behavior to take into account the gender biases of observers [20], and this intervention may be easier for them to implement than other behavioral adjustments (for instance, over-preparing for class or carefully curating ones appearance [20]) used to circumvent gender bias in the classroom.

Even though 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). While RQ1 measured a short-term effect (the survey was administered immediately after the grades were received), the intervention appears to have created a longer-term impact in student perceptions of the professor that carried through to the end-of-term evaluations. 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 [17]. Given that it is uncommon for students to receive detailed, positive feedback about their exam performance, the intervention email from the professor likely made a large impression on the students who received it.

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. The fact that ratings of the professor's likability were only increased for top-performing students–who received a more positive message than bottom-performing students–suggests that positive feedback is an essential component of the effectiveness of the intervention email. However, we cannot rule out the possibility that the email increased the professor's SETs because it provided additional attention from the professor and signaled concern for students, and that it was this aspect of the email that increased the professor's SETs. In other words, it is possible that bottom-performing students also gave the professor higher SETs after receiving the intervention, even if they did not rate her as more likable after receiving the email. Future research should investigate these nuances.

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 [7]. 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

There are factors other than the intervention that could have improved the professor's SETs during the Fall 2018 semester, and we explore these other possible factors here.

Were the SETs better during the intervention semester because the class size was smaller? One might argue that SETs were better during 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). Using 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, 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). This provides additional evidence that it was the intervention that increased SETs, not differences in class size between semesters.

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 (questions 3–6, Table III), which leads us to believe the delivery of material was similar between semesters. Additionally, the professor did not know which students opted in to the research, so no special treatment (or lack thereof) was given to students based on participation.

Were the course materials better during the intervention semester? One might argue that the intervention semester had better course materials. First, there were no observed differences in the SETs for questions related to the course material (questions 9–11, Table III). Second, 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 III-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 ($p = 0.184$). Thus, differences in SET response rates are unlikely to be the cause of higher SETs during the intervention semester.

B. Threats to Validity

All studies have threats to their validity; here we identify the most likely threats.

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 and replication is needed to assess the impact of response rate on the results of RQ2. However, student evaluations of teaching are reported for all students regardless of intervention or performance level. Even so, we observe significant differences between the two semesters, specially for questions related to the professor. This leads us to believe that the intervention had an effect.

3) Internal Validity: Communication among the students may have impacted the measured variables. As only half the students in the intervention semester received the positive, extra feedback on the exam scores, students in the control group who knew they received less performance feedback than their peers may have been more negative on the survey responses, leading to artificial differences in survey responses. However, our results show a short-term impact on student perceptions of professor likeability (RQ1) as well as a longer term impact on SETs (RQ2), leading us to believe the impact of this potential threat is minor.

VI. Conclusion

Research finds that SETs contain significant gender bias, and that professors who are women often receive lower evaluations 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 overall SETs, providing longer-term evidence of the intervention's effectiveness. While this intervention does not solve the gendered disparities that result from gender biases, this intervention could help women mitigate some of the bias they experience in SETs.

Acknowledgements

This work was supported in part by NSF #1749936.

References

[1] 2019. blinded. *blinded* (2019). under review.

[2] Alexandra C Achen and Paul N Courant. 2009. What are grades made of? *Journal of Economic Perspectives* 23, 3 (2009), 77–92.

[3] Emily T Amanatullah and Michael W Morris. 2010. 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* 98, 2 (2010), 256.

[4] American Sociological Association. 2019. Statement on Student Evaluations of Teaching. (2019). https://www.asanet.org/sites/default/files/asa_statement_on_student_evaluations_of_teaching_sept52019.pdf

[5] Susan A Basow and Nancy T Silberg. 1987. Student evaluations of college professors: Are female and male professors rated differently? *Journal of educational psychology* 79, 3 (1987), 308.

[6] Anne Boring, Kellie Ottoboni, and Philip Stark. 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research* (2016).

[7] Scott E Carrell, Marianne E Page, and James E West. 2010. Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics* 125, 3 (2010), 1101–1144.

[8] John A Centra and Noreen B Gaubatz. 2000. Is there gender bias in student evaluations of teaching? *The journal of higher education* 71, 1 (2000), 17–33.

[9] Alice Hendrickson Eagly, Linda L Carli Alice H Eagly, and Linda Lorene Carli. 2007. *Through the labyrinth: The truth about how women become leaders*. Harvard Business Press.

[10] Phil Edwards. 2015. Minions, Explained. (2015). https://www.vox.com/2015/7/10/8928069/minions

[11] Daniela Feistauer and Tobias Richter. 2018. Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. *Studies in Educational Evaluation* 59 (2018), 168–178.

[12] Diane Kierstead, Patti D'agostino, and Heidi Dill. 1988. Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology* 80, 3 (1988), 342.

[13] Lillian MacNell, Adam Driscoll, and Andrea N Hunt. 2015. What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education* 40, 4 (2015), 291–303.

[14] Kristina MW Mitchell and Jonathan Martin. 2018. Gender bias in student evaluations. *PS: Political Science & Politics* 51, 3 (2018), 648–652.

[15] Deborah A Prentice and Erica Carranza. 2002. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of women quarterly* 26, 4 (2002), 269–281.

[16] Kevin Rask. 2010. Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review* 29, 6 (2010), 892–900.

[17] Kelly G Shaver. 2016. *An introduction to attribution processes*. Routledge.

[18] Lisa Sinclair and Ziva Kunda. 2000. Motivated Stereotyping of Women: She's Fine if She Praised Me But Incompetent if She Criticized Me. *Personality and social psychology bulletin* 26, 11 (2000), 1329–1342.

[19] Anne Statham, Laurel Richardson, and Judith A Cook. 1991. *Gender and university teaching: A negotiated difference*. SUNY Press.

[20] Joan C Williams and Rachel Dempsey. 2018. *What works for women at work: Four patterns working women need to know*. NYU Press.

[21] S. Zweben and B. Bizot. 2018. The Taulbee Survey. *Computing Research Association* (2018). https://cra.org/resources/taulbee-survey/