Gender Bias in Student Evaluations

Kristina M.W.Mitchell, Ph.D.
Texas Tech University

Jonathan Martin, Ph.D.
Midland College

Under Review at PS: Political Science and Politics

> *"I want you personally to know I have hated every day in your course, and if I wasn't forced to take this . . . I never would have. Anytime you mention this course to anyone who has ever taken it, they automatically know that you are a horrific teacher, and that they will hate every day in your class. Be a human being show some sympathy everyone hates this class and the material so be realistic and work with people. "*
>
> -Excerpt from student email to a female online professor

Introduction

Are student evaluations of teaching (SET) biased against women, and what are the implications of this bias? While not unanimous in their findings, previous works have found evidence of gender bias in SET for both face-to-face and online courses. Specifically, evidence suggests that female instructors are rated lower than male instructors on SET due to gender. The literature examining gender bias in student evaluations is vast and growing (Bray and Howard 1980; Basow and Silber 1987; Miller and Chamberlin 2000), only more recently have scholars turned their focus on the potential of gender bias in the SET of online college courses. The use of online courses to measure gender bias offers a unique opportunity: the opportunity to hold constant many factors about a student's experience in a course that would vary in a face-to-face format.

The importance of SET vary from institution to institution as well as from position to position. Even with this variation, SET can influence decisions on hiring, tenure, raises, and

other employment decisions. Moreover, universities that place a high emphasis on the results of SET may be promoting discriminatory practices without recognizing it.

In this article, we look for sources of gender bias and argue that female instructors are evaluated differently from male instructors in two key ways. First, female instructors are evaluated based on a different set of criteria than male instructors, such as personality, appearance, or perceptions of qualification/competency. To test this, we use a novel method: a content analysis of student comments in official open-ended course evaluations and in online anonymous commentary. The evidence from the content analysis suggests that female instructors are evaluated more on personality and appearance, and are more likely to be labeled a 'teacher' instead of a 'professor.' Second, and perhaps more importantly, we argue that female instructors are rated more poorly than male instructors even in identical courses and when all personality, appearance, and other factors are held constant. We compare the SET of two instructors, one man and one woman, in identical online courses using the same assignments and course format. In this analysis, we find strong evidence to suggest gender bias in SET. The article concludes with comments on future avenues of research and on the future of the SET as a tool for evaluating faculty performance.

Gender Bias in Student Evaluations of Teachers

Measuring the impact of instructor gender on SET can be a difficult task due to the difficulty controlling for instructor-specific attributes. For instance, if a female instructor is evaluated poorly by students and a male instructor is evaluated highly, this could be due to gender bias or to instructor-related attributes such as teaching style or the overall quality of the teacher. In an effort to tackle this problem, MacNell et al. (2015) performed an experiment in which two assistant instructors (one male and one female) would each teach two sections of an

online course, one as a male instructor and another as a female instructor. With the ability to give the exact same course material to students and control for instructor quality, MacNell et al. found that students in the experiment gave the male identity a higher rating than the instructor with the female identity. Similarly, Boring (2017) in a similar research project used courses in which students were assigned at random to male or female professors and compared evaluations across sections, finding that female professors receive lower scores than male professors.

The empirical analysis presented in this article seeks to confirm these findings and the findings of others (Rosen 2017; Martin 2016). Specifically, our analysis both confirms the existence of bias in SET within the discipline of political science, while also contributing to the growing literature that suggests the problem of gender bias in SET is persistent throughout academia. If, as the mounting evidence suggests, SET are biased against women, then the use of evaluations in hiring is discriminatory.

Evaluation Criteria of Female Professionals

Gender bias is not only a question of whether male and female professors are evaluated more or less favorably. We argue that women are also judged on a different set of criteria than their men counterparts. We argue that women are evaluated differently in at least two ways: qualification/competence and personality.

In academia, female instructors are often viewed as not being as qualified compared to male instructors, and in many situations female instructors are perceived as having a lower academic rank than male instructors. One reason why women are stereotyped in this manner is because academia is still primarily a male dominated profession. As one example, women who serve as full-time employees are more likely to be in non-tenure track positions than men (Curtis 2011).

This qualifications stereotype is most evident in situations where a female professor is incorrectly given the lower rank of instructor by a student, and a male professor is accurately labeled as a professor. While not tested in the context of SET, Miller and Chamberlain (2000) find evidence to support the argument that women are more likely to be viewed as "teachers" whereas men are more likely to be referred to as "professors," indicating that students may view their female professor as not having as much experience or education, or being of lesser accomplishment than a male professor. If this qualification stereotype exists, it should be detectable in SET and student commentary. Thus, we hypothesize that male instructors are more likely to be referred to as "professors' in their SET, while female instructors will more likely be referred to as "teachers."

In addition to the qualifications stereotype, female instructors are more likely to be evaluated based on their personality. In previous studies on gender in academia, women have been viewed as having "warmer" personalities (Bennett 1982). In addition, Bennett (1982) also finds that students require female instructors to offer more interpersonal support than male instructors. In addition to being described as "warm," women have also been stereotyped as needing to exhibit nurturing and sensitive attitudes, such as being kind and sympathetic, to other people (Heilman and Okimoto 2007). Note the excerpt from a student email to an online female professor presented at the beginning of this article: the student specifically requests sympathy from the female professor.

Findings that female instructors are evaluated based on difference criteria than their male counterparts has broad implications for women in all professional fields, as these systematic differences may not be exclusive to academia.

Gender Bias in Student Comments & Reviews

To begin the exploration into gender bias in student evaluations, we first explore whether it is true that students evaluate female instructors using different criteria than male instructors. Our contribution is unique in its use of content analysis to examine student comments about their instructors from two sources. The first is the comments that students provided at the conclusion on their semester via official course evaluations. The second is comments uploaded to the popular site, Rate My Professors (RMP). RMP is a website in which students may anonymously post reviews of their instructors for other students to use when selecting a course or instructor. While RMP is not used for hiring or promotion decisions and, of course, suffers from selection bias, consistently poor evaluations in a public forum such as RMP can have career implications for academics, whether through lower enrollments in courses or an unfavorable reputation on campus.

If students consistently use different language to evaluate male and female professors, this language can set the tone for discussions about the quality of female instructors. Allowing students to provide open-ended comments about professors, whether via official evaluations or anonymous web evaluations, may create a platform for students to exhibit sexist tendencies, implying that these sexist comments matter. More importantly, it shows that even younger generations exhibit gender bias, and that sexism is not a relic of an older "good old boys" generation.

Each comment was analyzed for the following themes/topics: personality, appearance, entertainment, intelligence/competency, incompetency, referring to the instructor as "professor," and referring to the instructor as "teacher." A description of each theme, and examples of

comments within each theme, can be found in the appendix, as well as a more detailed explanation of the coding process.

We hypothesize that, regardless of the positive or negative nature of the overall commentary on each instructor, a female instructor will receive more comments that address her personality and appearance, fewer comments that refer to her as "professor," and more comments that refer to her as "teacher." We also hypothesize that a male instructor will receive more comments that discuss his intelligence or competency, fewer comments that refer to him as "teacher," and more comments that refer to him as "professor."

H_1: For categories Intelligence/Competency, Referred to as "professor"

$$Proportion_{Male} > Proportion_{Female}$$

H_2: For categories Personality, Appearance, Entertainment, Incompetency, Referred to as "teacher"

$$Proportion_{Male} < Proportion_{Female}$$

The percentage of comments that characterized each theme was compared. The results of the content analysis of official course evaluations for face-to-face courses are presented in Table 1.1, and the Rate-My Professors content analysis is presented in Table 1.2.

Table 1.1 – Content Analysis for Official University Course Evaluations

Theme	Professor Male	Professor Female	Difference
Personality	4.3%	15.6%	-11.2***
Appearance	0%	0%	0
Entertainment	15.2%	32.2%	-17***
Intelligence/Competency	13.0%	11.0%	2.0
Incompetency	0%	0%	0
Referred to as "professor"	32.7%	15.6%	17.1***
Referred to as "teacher"	15.2%	24.4%	-9.2**

N-68; *p<0.1; **p<0.05; ***p<0.01

The official student evaluations of face-to-face courses provided an interesting insight into the words students use to describe male instructors versus female instructors. The results

were generally as hypothesized, reflecting that students in official course evaluations do seem to use different language in evaluating male versus female instructions.

Table 1.2 – Content Analysis for Rate My Professors Comments

Theme	Professor Male	Professor Female	Difference
Personality	11.0%	20.9%	-9.9**
Appearance	0%	10.6%	-10.6**
Entertainment	5.5%	3.3%	2.3
Intelligence/Competency	0%	1.1%	-1.1
Incompetency	0%	6.6%	-6.6*
Referred to as "professor"	22.2%	22%	.3
Referred to as "teacher"	0%	5.5%	-5.5**

N=54; *p<0.1; **p<0.05; ***p<0.01

When the comments on the RMP website were analyzed, some of the differences between comments on a male versus female instructor were even more dramatic. Differences in mentions of appearance specifically became much more obvious in an anonymous forum.

Notably, the RMP comments on her personality tended to be negative (rude, unapproachable), while the official student evaluations tended to be more positive (nice, funny). In reading the context of each comment, the students mentioning her personality on RMP were almost exclusively taking an online course. Dr. Male taught an identical online course during the same time period as Dr. Female. The disparity between comments made by students taking identical online courses led us to another line of investigation: an empirical analysis of student ratings of identical online courses with a male versus female instructor.

Empirical Analysis of Gender Bias in Student Evaluations

In Spring 2015, both Dr. Female and Dr. Male acted as instructor of record for several online introductory political science courses. A detailed description of the courses and university

is provided in the appendix. We compared the ordinal evaluations of a male versus female instructor in five sections of the online courses.

Variance Across Sections

The advantage of using online courses to compare evaluations is that there were few aspects of the course that had an opportunity to vary across sections, regardless of the instructor. The online courses were essentially identical, with the exceptions of the course grader and the contact with the instructor.

First, each section had a different grader for written work. While the graders did attend the same training and have the same rubric, it is inevitable that some graders had stricter standards than others. Students may, thus, perceive one instructor or section differently based on the grader of the course. To determine whether differences in grading may have affected evaluations, we examined grade averages across sections. Table 1.3 shows grade averages for final grades, discussion posts, and short answer assignments for Dr. Male's sections and Dr. Female's sections.

Table 1.3 – Grading Averages in Online Courses

	Dr. Male Course Averages	Dr. Female Course Averages	Approximate Difference
Final Grades	75.23	79.30	-4%
Discussion Posts	67.97	73.09	-5%
Short Answers	65.60	67.74	-2%

While the differences were not dramatic, the data shows that grade averages for final grades, discussion posts, and short answers were lower in Dr. Male's courses than in Dr. Female's courses. If students give lower ratings in course evaluations due to lower grades or a

perception of stricter grading criteria, then Dr. Male (the male instructor) should have lower ratings on his course evaluations than Dr. Female.

Second, it is, of course, possible that one of the instructors had a more favorable demeanor in dealing with students, either via email or office hours. While it is difficult to convey tone in an email, it is possible that Dr. Male, for example, may have written emails in a kinder or more respectful tone than Dr. Female. Several examples of emails sent by Drs. Male and Female are provided in the appendix.

Evaluation Data

At the conclusion of the semester, students were asked to complete a 23-question evaluation of the course and instructor as a routine part of the university's procedure. Students rated their opinion on each question on a 5 point Likert Scale, with 1 being "Strongly Disagree" and 5 being "Strongly Agree." Participation in the evaluation process was voluntary.

We first categorized the questions asked into five question types. Our contribution is unique in that we separate any question that might have been directly related to instructor characteristics, such as sympathy or helpfulness, from those that did not vary across sections. The question categories were:

- **Instructor:** Questions specific to an individual instructor's characteristics, such as effectiveness, fairness, and encouragement.
- **Instructor/Course**: Questions that mentioned the instructor, but did not vary across the five sections, such as the instructor's ability to present information or stimulate learning.
- **Course**: Questions only about the course itself, including the expectations, workload, and experience.
- **Technology**: Questions that related to the technology in the course, such as information available related to technology.
- **Administrative**: Questions about registration, advising, or accessibility. The instructors had no ability to control or influence these factors.

Analysis

The ordinal ratings of each instructor in each section were averaged and compared. In the Instructor category of questions, a statistically significant difference would not necessarily indicate the existence of gender bias. Because these questions could have been influenced by personal characteristics, we present no hypothesis on the relationship between evaluations of the two instructors in this category.

If students exhibit a gender bias in their evaluation of their instructors, then Dr. Male will receive statistically significantly higher evaluations than Dr. Female in the categories of Instructor/Course, Course, and Technology. These questions relate to characteristics that are specific to the course, but do not vary across sections of the course. We predict that Dr. Male will receive higher evaluations than Dr. Female on questions in these three categories due to bias against female instructors.

Finally, because questions in the Administrative category address university-level issues that are not specific to an individual course or instructor, we expect to see no statistically significant difference between the evaluations of Drs. Male and Female in this category.

H_1: For categories Instructor/Course, Course, and Technology

$$Evaluation\ Average_{Male} > Evaluation\ Average_{Female}$$

H_2: For category Administrative

$$Evaluation\ Average_{Male} = Evaluation\ Average_{Female}$$

The results of the comparison were astounding. In every category except Administrative, Dr. Male received higher evaluations, including the non-instructor specific categories of Instructor/Course, Course, and Technology. An unpaired t-test was used to determine whether these differences were statistically significant. Table 1.4 shows the results of the unpaired t-test. Comparison results for each individual question are provided in the appendix.

Table 1.4 – Unpaired T-Test of SET by Category

	N	Mean Rating	Difference	T	P
Instructor					
Dr. Male	255	3.84	0.4***	5.24	0.000
Dr. Female	835	3.44			
Instructor/Course					
Dr. Male	255	3.71	0.4***	4.63	0.000
Dr. Female	835	3.31			
Course					
Dr. Male	357	3.71	0.22***	3.11	0.001
Dr. Female	1169	3.49			
Technology					
Dr. Male	153	3.83	0.19**	1.93	0.027
Dr. Female	501	3.64			
Administrative					
Dr. Male	153	3.96	-0.01	0.08	0.533
Dr. Female	501	3.97			

The data is clear in demonstrating that a male instructor received higher evaluations in identical courses, even for questions that have nothing to do with the individual instructor's ability, demeanor, or attitude.

However, perhaps it is possible that Dr. Female was so much worse as an instructor that students, in their ire, rated her lower in all categories, even those that had nothing to do with her or her course. This would be a valid critique, were it not for the final category of questions: the Administrative category. These questions asked students to evaluate university-level procedures, such as registration and advising. If students were simply unilaterally assigning low evaluations to Dr. Female without considering the question, then the questions in Administrative category would also have statistically lower ratings for Dr. Female.

On the contrary, in the Administrative category, there was virtually no difference in evaluations, with Dr. Female receiving an average rating .25% higher than Dr. Male. This was, as expected, not statistically significant.

It seems that students were considering the content of each question when responding. Even when no difference existed between Dr. Female and Dr. Male in course-level questions, students still rated the instructors either statistically similar, or they rated Dr. Female significantly less favorably than Dr. Male. To reiterate, *there were no questions out of the 23 questions asked in which a female instructor received a higher rating.*

Conclusion

Are student evaluations biased against women, and why does it matter? Our analysis of comments in both formal student evaluations and in informal online ratings indicate that students do evaluate female professors in a significantly different way than male professors. Students tend to comment on a female professor's appearance and personality far more often than a male professor. Female professors are also referred to as "teacher" far more often than male professors, which indicates that students may generally have less professional respect for their female professors. Based on our empirical evidence of online course evaluations, bias does not seem to be based solely (or even primarily) on teaching style or even grading patterns. Students appear to evaluate female professors poorly simply because they are female.

But more importantly is the question of why this matters. Many universities, colleges, and programs use student evaluations to make decisions on hiring, firing, and tenure. Because SETs are systematically biased against women, using them in personnel decisions is discriminatory. In addition, this could have broader implications for women in all professional fields.

Research in this field is far from complete. Our findings are a critical contribution, but more research will be needed before the longstanding tradition of using SET in employment

decisions can be eliminated. In addition, bias may not be limited to women. Our future research will examine not only gender bias in SET, but also bias in race, ethnicity, and English proficiency. Women have long claimed that their male counterparts are perceived as more competent and qualified. With mounting empirical evidence that this is true, perhaps it is time that universities use a method other than student evaluations to make these critical personnel decisions.

References

Basow, Susan A. and Nancy T. Silberg. 1987. "Student Evaluations of College Professors: Are Female and Male Professors Rated Differently?" *Journal of Educational Psychology* 79 (3): 308-314.

Bennett, Sheila K. 1982. "Student Perceptions of and Expectations for Male and Female Instructors: Evidence Relating to the Question of Gender Bias in Teaching Evaluation." *Journal of Educational Psychology*. 74 (2):170-179.

Boring, Anne. 2017. "Gender biases in student evaluations of teaching." Journal of Public Economics. 145: 27-41.

Bray, James H. and George S. Howard. 1980. "Interaction of Teacher and Student Sex and Sex Role Orientations and Student Evaluations of College Instruction." *Contemporary Educational Psychology* 5:241-248.

Correll, Shelley J., Stephen Benard, and In Paik. 2007. "Getting a Job: Is There a Motherhood Penalty?" *American Journal of Sociology* 112 (March):1297-1339.

Curtis, John W. 2011. "Persistent Inequity: Gender and Academic Employment." *Report from the American Association of University Professors*. Retrieved from

http://www.aaup.org/NR/rdonlyres/08E023ABE6D8-4DBD-99A0-24E5EB73A760/0/persistent_inequity.pdf

Cuddy, Amy J.C., Susan T. Fiske, and Peter Glick. 2004. “When Professionals Become Mothers, Warmth Doesn’t Cut the Ice.” *Journal of Social Issues* (60):701–718.

Elmore, Patricia B. and Karen A. LaPointe. 1974. “Effects of teacher sex and student sex on the evaluation of college instructors.” *Journal of Educational Psychology* 66 (June):386-389.

Heilman, Madeline E. and Tyler G. Okimoto. 2007. “Why Are Women Penalized for Success at Male Tasks?: The Implied Communality Deficit.” *Journal of Applied Psychology* 92(1):81-92.

Kaschak, Ellyn. 1978. “Sex Bias in Student Evaluations of College Professors.” *Psychology of Women Quarterly* 2(spring):235-243.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2015. “What’s in a Name: Exposing Gender Bias in Student Ratings of Teaching.” Journal of Collective Bargaining in the Academy. 0: Article 52.

Martin, Lisa. 2016. “Gender, Teaching Evaluations, and Professional Success in Political Science.” PS: Political Science and Politics 49(2): 313-19.

McKeachie, Wilbert J. 1979. “Student ratings of faculty: A reprise.” *Bulletin of the AAUP* 55:384-397.

Miller, JoAnn and Marilyn Chamberlin. 2000. “Women Are Teachers, Men Are Professors: A Study of Student Perceptions.” *Teaching Sociology* 28 (October):283-298.

Rosen, Andrew S. 2017. “Correlations, trends, and potential biases among publicly accessible web-based student evaluations of teaching.” Assessment & Evaluation in Higher Education.