

Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity

Kerry Chávez, *Texas Tech University*

Kristina M.W. Mitchell, *San Jose State University*

ABSTRACT Research continues to accumulate showing that in instructor evaluations students are biased against women. This article extends these analyses by examining the dynamics between evaluations and gender and race/ethnicity. In a quasi-experimental design, faculty members teaching identical online courses recorded welcome videos that were presented to students at the course onset, constituting the sole exposure to perceived gender and race/ethnicity. This enables exploration of whether and to what degree the instructors' characteristics influenced student evaluations, even after holding all other course factors constant. Findings show that instructors who are female and persons of color receive lower scores on ordinal student evaluations than those who are white males. Overall, we add further evidence to a growing literature calling for student evaluations of teaching (SETs) reform and extend it to encompass the effects on racial/ethnic minorities in addition to women.

The use of student evaluations of teaching (SETs) is widespread in colleges across the country for career decisions including hiring, tenure, and promotion. Mounting research demonstrates that SETs are biased against women, potentially resulting in discriminatory personnel practices to the degree that they are weighted in such decisions. Based on theories underlying gender bias, we suspect that SETs also are biased against other marginal groups including racial/ethnic minorities. Using a quasi-experimental research design, we examined SET results for 14 online courses designed to be identical except for the identity of the instructors. Even with all course elements held constant—content, assignments, schedules, and communications—women and faculty of color received lower scores than white men. This is consistent with many other findings on bias in academia and with other quasi- and non-experimental works, further confirming that SETs exhibit bias in favor of white male instructors. Our study adds to this growing corpus by offering yet another instance of evidence of bias and extending the framework to include race/ethnicity.


THE ALREADY COMPELLING LITERATURE


Scholars began heeding potential gender bias in SETs in the 1980s, yielding mixed results that impelled little motivation

or guidance for addressing it. The ambiguous findings abraded against conventional wisdom and anecdotal experiences. Laube et al. (2007, 90) commented:

Frankly, as sociologists specializing in gender, we are puzzled by conclusions that gender has no impact on teaching evaluations. Three decades of scholarship have found gender to be a significant factor in shaping interactions, practices, and outcomes in every major realm of human social life.... Why would the classroom be any different?

Thus, suspicions of bias lingered, leading to better research designs and greater rigor in observational and experimental analyses that began to show consistent evidence of its operation. The recent wave of scholarship confirms evidence of both direct and indirect gender bias in academic evaluative processes. In its direct form, students rate female instructors lower than males due to nothing other than gender. Using final exams as an independent measure of student learning, Boring (2017) demonstrated that female instructors receive lower scores despite objective indications that they are as equally effective and efficient as their male colleagues. Miller and Chamberlin (2000) discovered that students tend to appellate women as teachers and men as professors regardless of their actual positions or credentials.¹ MacNell, Driscoll, and Hunt (2015) found that students rate those *perceived* to be female (even when they are male in reality) lower in an online class experiment in all categories, more than half of them being statistically significant. In summary, all else being equal, students perceive and evaluate female instructors more critically.

Kerry Chávez  is a PhD candidate in political science at Texas Tech University. She can be reached at kerry.chavez@ttu.edu.

Kristina M.W. Mitchell  is a lecturer at San Jose State University. She can be reached at kristina.mitchell@sjsu.edu.

All else is not equal. In addition to direct bias, studies have uncovered evidence of indirect bias implying that students evaluate women on different and additional criteria than men (Boring 2017; Mitchell and Martin 2018). Men usually are perceived as agentic types, being more assertive, ambitious, and independent. Women are categorized as communal types, expected to exhibit helpfulness, sensitivity, and kindness (Eagly and Karau 2002; Martin 2016). Based on role-(in)congruity theory, empirical support is amassing that students employ stereotypes in the evaluative process that generate cross-pressures on female instructors to fulfill requisites for both their gender and their professional role. Furthermore, they are evaluated more closely than males, suggesting that students do not tolerate deviation from stereotypy. This constitutes a “double-bind dilemma” in which gendered expectations conflict between femaleness and higher education

them frequently tapped stereotypes. Based on role-incongruity theory and the notion of marginality, we expect that scholars of color receive systematically lower evaluation scores, thereby stunting their competitiveness with colleagues born more conveniently into students’ stereotypes. Just as gendered evaluations operate on “shifting standards” where one making a judgment is compelled to do so relative to a reference point, we assert that the same process occurs with people of color and accent in comparison to white males with native linguistic inflections. As a result, we submit that scholars of minority race and ethnicity contend with similar structural barriers to success and status, traveling their own “leaky pipeline” that has yet to be surveyed. The next section presents evidence of this that we hope spurs further research to deconstruct the determinants of racial bias in SETs.

Just as gendered evaluations operate on “shifting standards” where one making a judgment is compelled to do so relative to a reference point, we assert that the same process occurs with people of color and accent in comparison to white males with native linguistic inflections.

(MacNell, Driscoll, and Hunt 2015). This is exacerbated in the context of leadership, which closely aligns with the role of an academic instructor (Johnson et al. 2008; Martin 2016). Moreover, both direct and indirect bias are exacerbated in fields perceived to be masculine in nature or exhibiting steeper underrepresentation of women (Eagly and Karau 2002).²

So what? In the first place, Curtis (2011, 5) acknowledged that “even after four decades of focused attention and policy development around the issue of gender equity in academia, women have not achieved the same status as men.” Females are less likely to be full-time tenure-track, to hold tenured positions, to attain higher leadership roles in academia, and to earn the same salary as males in the same positions (Curtis 2011). Women also are underrepresented in publications (especially top-tier journals), receive differential returns on investment in coauthorships, and are cited disproportionately less by men—all assemblages of evidence that the discipline reproduces these biases (Djupe, Smith, and Sokhey 2019; Mitchell, Lange, and Brus 2013; Teele and Thelen 2017). Rather than a single hurdle inhibiting equitable outcomes, there are gendered disadvantages strewn throughout most of the career trajectory. We argue that SETs might be one of them. Although the importance of SETs varies among institutions, they often are used as an indicator of teaching effectiveness, for hiring and tenure decisions, and in considerations of raises and promotions (Laube et al. 2007; Mitchell and Martin 2018). Quantitative evaluations are being implemented more frequently and weighted more heavily, likely because they constitute a low-cost metric of performance that automatically aggregates and interprets responses. Insofar as this impels administrators toward SETs, and insofar as quantitative formats assume that students are applying the same criteria across genders in a way that obscures perceptual bias (Laube et al. 2007), they might be more problematic.

Although the majority of existing research revolves around gender bias, we posit that potential direct and indirect biases against instructors of other races and ethnicities are likely as well. Gender, race, and ethnicity are highly visible, making

YET ANOTHER COMPELLING RESEARCH DESIGN

In the summer of 2017, 14 online political science sections of two courses—Introduction to American Government and Introduction to Texas Government—were offered at a large state institution. Students took the courses to fulfill a mandatory component of a bachelor’s degree at the university. A different faculty member presided over each section as the instructor of record to supervise grading and broad course-level issues. Although the instructors were considerably visible—listed in the registration interface, on the online course homepage, as the suggested first point of contact for questions or concerns, and featured in a welcome video posted to the course homepage and emailed to all relevant students—their interaction with students ended there.

Each American government course contained identical textbook readings, lecture content, activities, assignments, and student assessments. The Texas government courses also were sourced and structured identically. All welcome videos recorded by the instructors contained an identical script that delivered a brief and basic welcome to the course. Most innovatively, all 14 courses had the same course coordinator charged with overseeing day-to-day administrative responsibilities and to ensure consistency across the sections. In the summer of 2017, instructors made use of the course coordinator to handle all questions and communications from students. Although all students had a different professor, their interactions remained with the course coordinator, who responded to students across courses with high consistency.³ In essence, every component of the online courses was identical except for the identity of the professor. This constituted an ideal quasi-experimental situation in which to compare SETs to determine the impact of gender and race/ethnicity on evaluations.

Student Perceptions

Whereas gender often is easy to perceive based on an individual’s name and welcome video, a professor’s racial or ethnic identity might be different than a student’s perception. Because it is the students who are evaluating their professors, it is their perceptions

of race/ethnicity that matter. For this reason, a survey was conducted to tap students' perceptions of the course professors. The 48 students canvased were from the same university but were not enrolled in the summer online courses. In a sophomore-level philosophy course, respondents were assigned to view still photographs taken from the instructors' welcome videos and to respond to a series of questions probing their perceptions. Unsurprisingly, students were universally correct in identifying professors by their self-articulated genders. Students were generally accurate in identifying professors as white or nonwhite and in selecting the correct racial/ethnic identity, although less so than with gender. For example, a white Hispanic professor was identified by 87% of student respondents as white non-Hispanic. For the purposes of the survey, faculty members were coded as nonwhite if at least

We recorded five students providing positive mentions about their white professors and two submitting negative mentions. Finally, one student wrote a positive comment and one a negative note about their nonwhite professor. We found no evidence of students explicitly linking their distaste or praise to gender or race. However, given that there was no variation across course sections other than instructor identity, it is implicitly concerning that there is a noticeable difference between the types of comments that women and faculty of color received. We consider this tabulation, summarized in table 1, as preliminary evidence of bias—although, admittedly, the sample size reduces confidence and reliability of results. To overcome this and substantiate the patterns, the next section is a more rigorous analysis of student evaluations.

“She got super annoyed when people would email her and did not come off very approachable or helpful.”

60% of respondents identified them as such—a threshold that captures a majority perspective but permits greater variation because race/ethnicity is not always visually overt.

Student Evaluations

At the conclusion of the summer term, students were requested to complete course evaluations composed of nine questions⁴ related to the quality of the instructor, the course, and their university-level experiences.⁵ Following the official prompts, respondents were given the opportunity to fill out an open-ended section with additional comments. The next section is a qualitative analysis of the relevant remarks provided in the comment section and statistical analysis of the ordinal evaluations.

THE DISCONCERTING RESULTS

Across the qualitative and quantitative approaches, we find consistent evidence of bias at work in student evaluations of women and to a lesser degree of instructors of nonwhite race/ethnicity.

Student Evaluation Commentary

The majority of comments provided by students were appropriate and course-related. For example, many students had positive reviews of the course content or organization and many had complaints about the platform or learning management system. Some students, however, took the opportunity in the open-ended prompt to comment on professors. Because professors did not interact with students outside of the welcome video, the disparity between comments on male versus female or white versus nonwhite professors can be attributed to underlying bias. All instructor-specific comments were identified and categorized as positive or negative. For example, “Most intelligent, well-spoken, cooperative professor I have had the chance to encounter”⁶ was considered positive whereas “She got super annoyed when people would email her and did not come off very approachable or helpful”⁷ was coded as negative.

Four students volunteered positive comments in their evaluations of their male professors; none submitted negative commentary for this category. Two students made positive comments of female professors, but three made negative remarks about them.

Statistical Analysis

In addition to the open-ended comments, students were asked nine questions at the conclusion of the semester to evaluate their professor, course, and university experience. Three questions focused on the student's experience with the instructor. Because the impact of instructor-specific characteristics was the subject of analysis, we relied primarily on these questions for comparison. As mentioned previously, there were 14 instructors: 11 men and three women, eight white and six nonwhite.⁸ Although we would want a more balanced breakdown of gender, it has already been established that female faculty are underrepresented in political science. Because discerning the visual patterns between male and female or between white and nonwhite faculty is difficult from descriptive data, we relegate the listing of average scores to the online appendix. A t-test can better determine statistically significant differences between student evaluations across race and gender. At first glance, results presented in table 2 demonstrate that women received statistically significant lower evaluations, but nonwhite faculty members were not evaluated more harshly. One possible explanation is that stereotypes for the range of race/ethnicity are less cohesive than for the focused female category. Although minority faculty are marginal in certain respects, there is less theory to guide predictions on preconceptions or expectations that students bring to evaluations of various races/ethnicities and that we urge future work to address.

Table 1
Tabulation of Open-Ended Student Comments on Professors

	Positive Comments	Negative Comments
Men	4	0
Women	2	3
White	5	2
Nonwhite	1	1

We proceed with a more thorough analysis of the predictors of SETs before concluding whether gender and race matter in this quasi-experiment. A simple dataset was generated to facilitate regression analysis. The dependent variable is the average score for each professor on the three instructor-specific questions.

we relaxed our expectations for robustness given the sample size. Overall, we interpret this as weak initial evidence that similar patterns of bias are evident during assessments of instructors of different race and ethnicity. Because the evidence is less stark and the causal mechanisms are less developed, we call for further

Thus, in their current form, SETs might constitute another “weep hole” for women and minorities in academic career pipelines that structurally contribute to higher attrition and lower achievement.

The explanatory variables include gender (coded 0 for male, 1 for female) and race/ethnicity (coded 0 for white, 1 for nonwhite). Ideally, we would have run an interaction between female and nonwhite faculty, but constraints in our sample size inhibited this opportunity. One important control is included—the average final grade for each section of the course.⁹ Although every course in the summer term was identical in content and assignments, there was opportunity for variation in grades across sections because of either student characteristics or grading patterns. Results of the regression analysis (table 3) show that, controlling for final grades, gender and race remain significant predictors of SET scores. Substantively, using predicted probabilities, females received a 5.81% lower score by virtue of their gender and non-white instructors received a 3.94% lower score than their white colleagues. The slighter impact of race might explain why this variable fell short of statistical significance in table 2; however,

research on how color, physiognomy, and accent affect student perceptions and evaluations.

CONCLUSION

We recognize the limitations of the research results presented here. The sample size is small, the courses were limited to a single institutional setting that may be more or less likely to exhibit bias against women and minorities, and results are unable to capture student-specific factors other than grades that might affect evaluations. Nonetheless, the quasi-experimental design—in which all aspects of the course (except brief exposure to instructor identity) were held constant—creates an ideal situation in which to compare SETs. Our findings that students are biased against females and non-natives in their evaluations of teaching comports with numerous studies revealing the same phenomenon. Thus, in their current form, SETs might constitute another “weep hole” for women and minorities in academic career pipelines that structurally contribute to higher attrition and lower achievement. We agree with Martin (2016, 318) that rather than encouraging women and other minorities to “lean in and perform better within the current system,” it is high time academia brings its ingenuity to bear to develop better measures of teaching effectiveness.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit <https://doi.org/10.1017/S1049096519001744>. ■

Table 2

T-Tests of Differences Between Student Evaluations Across Gender and Race/Ethnicity

	Average Evaluation Score: Women	Average Evaluation Score: Men	Difference
Overall	4.0	4.2	0.2*
Instructor-Specific	4.1	4.3	0.2**
	Average Evaluation Score: Nonwhite	Average Evaluation Score: White	Difference
Overall	4.2	4.2	0.0
Instructor-Specific	4.2	4.3	-0.1

*0.1 significance level, **0.05 significance.

Table 3

OLS Regression Results of SETs from the Online Courses

	Model 1	Model 2
Gender	-0.23*** (0.09)	-0.24*** (0.09)
Minority	-0.14* (0.07)	-0.15* (0.08)
Final Grade		-0.009 (0.015)
N	42	42
R ²	0.21	0.22

*0.1 significance level, **0.05 significance, ***0.01 significance.

NOTES

1. I, Dr. Mitchell, once received an email from a student in a course cotaught with a male professor with the greeting “Dr. [male instructor’s surname] and Miss Kristina.”
2. Paludi and Strayer (1985) selected political science to represent the masculine discipline in their experiment on how students rate academic articles by name and field. Similarly, whereas Rosen (2018) found small statistically significant differences between male and female groups on RateMyProfessor.com, the gap is larger in certain fields, singling out political science as a paragon of the problem.
3. Although the coordinator was listed in a master directory at the bottom of the course homepage, which enabled students to ascertain gender and race/ethnicity, students believed they were communicating directly with their own listed professors. They had no reason to suspect that responses came from another individual and no resources to trace them back to the coordinator. Consequently, we do not believe that the demographics of the coordinator introduced bias into the research design.

4. The text and categorization of all questions are presented in the online appendix.
5. The overall response rate was 35%. We performed calculations to ensure that response rates for professors across our demographics of interest were not significantly different.
6. Said of a white male instructor.
7. Said of a white female instructor.
8. As coded from student perceptions.
9. Previous literature suggests that grades are influential in determining how students evaluate a course and a professor (Boring 2017; Laube et al. 2007; Rosen 2018).

REFERENCES

- Boring, Anne. 2017. "Gender Biases in Student Evaluations of Teaching." *Journal of Public Economics* 145: 27–41.
- Curtis, John W. 2011. "Persistent Inequity: Gender and Academic Employment." Report from the *American Association of University Professors*. Available at www.aaup.org/NR/rdonlyres/08E023AB-E6D8-4DBD-99A0-24E5EB73A760/0/persistent_inequity.pdf.
- Djupe, Paul A., Amy Erica Smith, and Anand Edward Sokhey. 2019. "Explaining Gender in the Journals: How Submission Practices Affect Publication Patterns in Political Science." *PS: Political Science & Politics* 52 (1): 71–77.
- Eagly, Alice H., and Steven J. Karau. 2002. "Role-Congruity Theory of Prejudice Toward Female Leaders." *Psychological Review* 109 (3): 573–98.
- Johnson, Stefanie K., Susan Elaine Murphy, Selamawit Zewdie, and Rebecca J. Reichard. 2008. "The Strong, Sensitive Type: Effects of Gender Stereotypes and Leadership Prototypes on the Evaluation of Male and Female Leaders." *Organizational Behavior and Human Decision Processes* 106 (1): 39–60.
- Laube, Heather, Kelley Massoni, Joey Sprague, and Abby L. Ferber. 2007. "The Impact of Gender on the Evaluation of Teaching: What We Know and What We Can Do." *National Women's Studies Association Journal* 19 (3): 87–104.
- MacNell, Lillian, 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): 291–303.
- Martin, Lisa L. 2016. "Gender, Teaching Evaluations, and Professional Success in Political Science." *PS: Political Science & Politics* 49 (2): 313–19.
- Miller, JoAnn, and Marilyn Chamberlin. 2000. "Women Are Teachers, Men are Professors: A Study of Student Perceptions." *Teaching Sociology* 28 (4): 283–98.
- Mitchell, Kristina M. W., and Jonathan Martin. 2018. "Gender Bias in Student Evaluations." *PS: Political Science & Politics* 51 (3): 1–5.
- Mitchell, Sara McLaughlin, Samantha Lange, and Holly Brus. 2013. "Gendered Citation Patterns in International Relations Journals." *International Studies Perspective* 14 (4): 485–92.
- Paludi, Michele A., and Lisa A. Strayer. 1985. "What's in an Author's Name? Differential Evaluations of Performance as a Function of Author's Name." *Sex Roles* 12 (3/4): 353–61.
- Rosen, Andrew S. 2018. "Correlations, Trends and Potential Biases Among Publicly Accessible Web-Based Student Evaluations of Teaching: A Large-Scale Study of RateMyProfessors.com Data." *Assessment & Evaluation in Higher Education* 43 (1): 31–44.
- Teele, Dawn Langan, and Kathleen Thelen. 2017. "Gender in the Journals: Publication Patterns in Political Science." *PS: Political Science & Politics* 50 (2): 433–77.