Gender, Teaching Evaluations, and Professional Success in Political Science

Lisa L. Martin, *University of Wisconsin, Madison*

ABSTRACT Evaluations of teaching effectiveness rely heavily on student evaluations of teaching. However, an accumulating body of evidence shows that these evaluations are subject to gender bias. Theories of leadership and role incongruity suggest that this bias should be especially prominent in large courses. This article examines publicly available data from two large political science departments and finds that female instructors receive substantively and significantly lower ratings than male instructors in large courses. The author discusses the implications of apparent gender bias in teaching evaluations for the professional success of female faculty. Findings of gender bias in evaluations in other fields also hold in political science and are particularly problematic in the evaluation of large courses.

Decisions about promotion and tenure in political science departments include an evaluation of teaching effectiveness. Although some universities have moved beyond sole reliance on student evaluations of teaching (SETs), they remain a core part of the teaching dossier. Many female faculty members believe that they face prejudice in SETs. However, skepticism remains about the existence or degree of gender bias in SETs. Historically, systematic studies of SETs were mixed in their findings of gender bias; however, newer and more rigorous studies show an emerging consensus that gender bias does exist. This article builds on the broad body of work on gender bias in SETs to extend these findings to political science departments and to introduce a new argument about the interaction between instructor gender and class size.

This article presents a number of interrelated arguments. Increasingly, the literature suggests that female instructors receive lower rankings than male instructors across a range of disciplines. In a twist on this research, I argue that the effect of an instructor's gender should be dependent on the size of the course. My review of the literature on gender and leadership assessments suggests that there should be an interaction between the gender of the instructor and student assumptions about leadership roles. Thus, when a course requires that a teacher take on a stereotypical leader role—such as a large lecture course—assumptions about gender roles could have a significant impact on evaluations. I provide an empirical assessment of the hypothesis about an interaction between class size and gender bias using publicly available SET data from two political science departments at large public universities. These data show, as expected, that female faculty members receive lower evaluations of general teaching effectiveness in large courses than male faculty members, whereas there is no substantial difference for small courses. To the extent that teaching evaluations are an important part of promotion and compensation decisions and other reward systems within universities, reliance on SETs that appear to be biased creates concerns. These concerns suggest that the discipline must reconsider its methods of faculty evaluations and the role that they have in professional advancement.

The first section of the article discusses the general literature on gender bias in SETs. The second section turns to theory, arguing that role-incongruity theory strongly indicates that there should be an interaction between the degree of gender bias and class size. The third section presents empirical evidence from two political science departments and concludes by drawing implications for the use of SETs in processes of professional advancement and reward.

GENDER BIAS IN EVALUATION OF TEACHING EFFECTIVENESS

The potential for gender bias in SETs has long been recognized and discussed. This section summarizes the general literature on gender and SETs and the more limited work on this relationship in the political science discipline. The role of class size is rarely

Lisa L. Martin *is professor of political science at the University of Wisconsin, Madison. She can be reached at lisa.martin@wisc.edu.*

doi:10.1017/S1049096516000275

mentioned in these studies. It is worth noting, first, that studies of possible gender bias in SETs in higher education began appearing in the 1980s and 1990s, and early findings were mixed (e.g., Basow and Silborg 1987; Centra and Gaubatz 2000; Feldman 1993; Sidanius and Crane 1989).

However, recent and more rigorous studies show consistent evidence of bias. These studies are based on both experiments and observational analysis. Arbuckle and Williams (2003) undertook a fascinating experiment in which students viewed a stick figure that delivered a short lecture. All participants observed the same stick figure and the same lecture but the figures were given labels of old or young and male or female. Participants significantly rated the figure labeled as a young male as the most expressive, which illustrates that students' expectations influence their perception of an instructor independent of the material or how it is delivered. A similar experimental setup in a distance-education course allowed researchers to manipulate whether a male or female instructor was teaching the course and whether students believed that the instructor was male or female (MacNell, Driscoll, and Hunt 2014). The authors found that "the male identity received significantly higher scores on professionalism, promptness, fairness, respectfulness, enthusiasm, giving praise, and the *student ratings index*" (MacNell, Driscoll, and Hunt 2014, 8), regardless of whether the instructor was actually male or female.

One particularly striking finding in this study was that even relatively objective questions, such as whether the instructor was prompt, led students to score the instructor almost one point lower on a five-point scale if they believed that the instructor was female. This finding suggests that the fault of SETs is not in the way that questions are posed or which qualities they ask about; rather, the fault lies in the nature of the instrument itself.

Other recent work relies on observational rather than experimental techniques. Miller and Chamberlin (2000) focused on students' perception of instructor educational credentials and found that they perceive male instructors as having higher or superior credentials. In a recent study undertaken in an Italian engineering college, Bianchini, Lissoni, and Pezzoni (2012) found that in three of the four programs they examined, women consistently received significantly lower effectiveness scores than men. The authors speculated that the gender composition of the student body could account for their findings because two of the four programs had low percentages of female students.

In an especially well-designed observational study, Boring (2015) compiled more than 22,000 observations of student ratings in a French school of social science. She examined mandatory introductory classes in which students' ability to choose their instructor is tightly constrained. The courses include a standard final examination that is graded anonymously, which provides an independent, objective measure of student learning. The numerous observations allowed Boring to control for both student and teacher fixed effects. All of these factors allowed for an unbiased and reliable measure of bias, representing a major improvement on other observational studies. They allowed Boring to not only measure the degree of gender bias in SETs but also to explore its roots and whether instructor ratings are a good indicator of teaching effectiveness.

Boring's results are striking. She found that male instructors receive significantly higher ratings, which results from a strong male-student bias. Male students are 30% more likely to give a rating of "excellent" to male than female teachers (Boring 2015, 5). Female instructors scored relatively well in more time-consuming tasks, such as course preparation, whereas male instructors scored well in less time-consuming activities, such as leadership skills. Boring also found that students who receive higher grades give higher instructor ratings, and she calculated that women could receive the same rating as men if they gave students a 7.5% boost in their grades (Boring 2015, 2). Because Boring used the final exam as an independent measure of student learning, she could explore the degree to which student performance is correlated with higher teacher ratings. She found that it is not correlated and that "SET scores do not seem to measure actual teaching effectiveness" (Boring 2015, 2).

Within political science, the APSA has occasionally published a piece in *PS* that draws attention to the potential for bias in SETs, and it offers advice for concerned faculty. Langbein (1994) noted that the effect of low grades on teaching evaluations is more pronounced for female than male faculty. Noting that poor evaluations can have negative effects on promotion and compensation decisions, Langbein questioned whether SETs are adequately valid measures of teaching effectiveness to have such an important role. Andersen and Miller (1997) noted that female instructors who are not perceived as caring and accessible may fail to meet student expectations and therefore may be penalized on SETs. Sampaio (2006) examined the intersection of gender, race, and subject matter, focusing on implications for women of color in the classroom. Dion (2008) reviewed the literature on bias and offered advice for women faculty who must be both authoritative and nurturing. In related work, Baldwin and Blattner (2003) suggested that because SETs may be biased, alternative evaluation measures should be considered. Smith (2012) noted that SETs are used for both professional development and employment decisions, setting up tensions. These tensions are especially pronounced, given questions about the validity and reliability of SETs as well as peer observation of teaching.

Small seminars allow for extensive one-on-one interaction and the ability to establish empathy while still demonstrating mastery of the material. However, in large lecture courses, the opportunities to exhibit sensitivity to individual students are more limited.

ROLE INCONGRUITY AND LEADERSHIP IN LARGE CLASSES

We can make more sense of studies of gender bias in SETs by turning to the psychology literature on role incongruity and leadership. A body of work known as "role-congruity theory" puts these studies of SETs in context and suggests more refined ways to approach the question of gender bias. The idea behind role-congruity theory is that individuals enter social interactions

with implicit assumptions about the roles that others will play. Gender roles are prominent in this literature, with men implicitly associated with the "agentic" type: more assertive, ambitious, and authoritative. Women tend to be implicitly associated with the non-agentic type: more passive, nurturing, and sensitive. Role incongruity occurs when a man or a woman acts in a way that is contrary to type—for example, if a woman takes on an agentic demeanor. A situation that demands that a woman be agentic will cause role incongruity and can lead to negative reactions from students. I link this body of theory to SETs by noting that some class settings demand a more agentic approach than others. Small seminars allow for extensive one-on-one interaction and the ability to establish empathy while still demonstrating mastery of the material. However, in large lecture courses, the opportunities to exhibit sensitivity to individual students are more limited. At the same time, these "sage-on-a-stage" formats demand that the instructor be assertive and demonstrate consistent authority.

Although the literature on role congruity and leadership is extensive, I summarize the studies linked most directly to my focus on SETs. Butler and Geis (1990) used experimental approaches to examine the role of gender and leadership in the reactions of observers. They focused on nonverbal responses—in particular, positive or negative facial reactions of participants who observed leaders making suggestions for certain courses of action. Female leaders elicited significantly more negative facial expressions than males in the same situation. Ridgeway (2001) discussed "gender status beliefs" and how they constrain individuals' expectations of leaders. Gender status beliefs lead individuals to assume that men will be more competent and assertive as leaders. Experiments that test these ideas reveal that when women are placed in a leadership role and act assertively, they are punished. Rudman and Glick (2001) also examined the potential for backlash against agentic women. They found that women who violate stereotypes by exhibiting intelligence, ambition, and assertiveness elicit negative reactions. However, this effect can be mitigated if women "temper their agency with niceness" (Rudman and Glick 2001, 743).

In Eagly and Karau's (2002) review of the work on role-congruity theory and female leadership, they found that two forms of prejudice are most prominent. First, women are generally viewed less favorably as leaders. Second, when women exhibit behaviors that are associated with leadership (e.g., projecting authority), they are evaluated less favorably than men. In a novel multimethod approach, Johnson et al. (2008) conducted a series of tests of role-congruity theory using qualitative, experimental, and survey approaches. They contrasted the "strong" (agentic) type to the "sensitive" (non-agentic) type. Consistent with other studies, they found that female leaders must project both strength and sensitivity to be effective, whereas male leaders need only project strength.

Taken as a whole, these studies argue for a more nuanced approach to the potential for gender bias in SETs. Different types of courses demand that instructors assume different roles. In small classes (e.g., seminars), the instructors usually are seated and their role is to guide discussion and draw out students' thoughts, thereby facilitating class discussion. In this setting, students likely do not come to class with expectations that the instructor will play the typical agentic-leader role. However, when contrasted to a large lecture course, when the instructor is on a stage with a microphone speaking in front of hundreds of students, the opportunities for interaction with individual students, to express concern for their specific needs, and to draw out their opinions are limited. Instead, students are likely to come to class with standard expectations of agentic leadership.

If this is the case, the potential for backlash against agentic women will be significant in large lecture settings, whereas it is likely to be minimal or absent in small class settings. Ratings for female instructors tend to decline with class size at a higher rate than for male instructors. This logic leads to the following hypothesis.

> Hypothesis 1: The interactive effect between male gender and class size on SETs will be positive.

Hypothesis 1 can explain why early studies did not find gender bias in SETs. Perhaps these biases primarily arise when leadership expectations are invoked—that is, in large classes. If women tend disproportionately to teach smaller classes than men (perhaps because of negative feedback when they attempt large courses), the interaction between course size and instructor gender could lead to average effects of gender being washed out. If this hypothesis is correct, then we need an interaction effect between class size and lower effectiveness ratings for female faculty in order to test it. The presence of such an effect would validate the relevance of role-congruity theory to the classroom and renew concerns about reliance on SETs as measures of teaching effectiveness.

Whereas other types of interaction effects between gender and other course characteristics have received attention, this specific interaction between course size and instructor gender has not been studied in depth. One exception is Wigington, Tollefson, and Rodriguez (1989), who collected data involving 5,843 student evaluations at a midwestern university in the mid-1980s. The authors found that the expected effect did appear: "The interaction between sex and size was due to males having higher ratings than females in the larger classes…" (Wigington, Tollefson, and Rodriguez 1989, 339). This effect was reversed for small classes. Unfortunately, the authors did not pursue this result any further and it apparently has gotten lost in a general sense that "interactions matter." More recently, in a study at a college of engineering, Johnson, Narayanan, and Sawaya (2013) found that female instructors receive lower ratings, as do larger classes. However, they did not examine the interaction between these two factors. The next section presents new evidence on the interaction between course size and instructor gender using data from political science departments.

EVIDENCE AND IMPLICATIONS

Today, only a few public universities make SET results publicly available. The following analysis is based on records from two political science departments in large, public research universities. One is a southern university, for which I have data from 2011 through 2014; the other is a western university, which includes data from 2007 through 2013. Total enrollment in the southern university is more than 58,000 and it is more than 31,000 in the western university. Both are well-ranked R1 research universities with large political science departments. Both administer their evaluations online. I collected all evaluations from undergraduate courses taught by faculty during the years indicated. According to the universities' own documentation, these evaluations are

required for consideration during promotion and tenure reviews. The southern university requires that the tenure dossier include a "complete longitudinal summary" of SETs in tabular form. The western university's guidelines are less precise but specify that SETs must be included as one of two forms of teaching evaluation. Therefore, these instruments have a direct impact on professional advancement at the two institutions.

To investigate the predicted interactive effect of gender and class size, I used Tobit analysis. This approach is appropriate because the data are censored at both the top and the bottom of the five-point scale. That is, even students who loved the class cannot give a score above five and those who hated it cannot give a score below one. Table 1 shows the results of Tobit analysis, examining the effect of gender, course size, and interaction between the two on average course evaluations.

For the southern university, the dependent variable in this analysis is the average response, on a five-point scale, to the statement: "Overall, this instructor was effective." "Strongly agree" is equivalent to five points and "strongly disagree" is equivalent to one point. Analysis is based on all 309 faculty evaluations available on the university's website for this time frame. Enrollment in courses was not available, so course size is estimated by the number of students who completed the evaluation.[1] The western university also uses a five-point scale. The question asked is whether students "learned from the course." Enrollment data are available for this university, and the dataset includes 587 evaluated courses.

For both universities, the evidence supports Hypothesis 1. The coefficients are in the expected direction, showing a positive interaction effect between a male instructor and a larger class. The results for the southern and western universities are statistically significant at the 0.10 and 0.05 levels, respectively. Table 2 and figure 1 summarize the estimated substantive effects.

For a small course with 10 students, there is little difference in ratings between male and female instructors. For a larger course with 100 students, a more sizeable difference emerges, with males scoring two tenths and one tenth of a point higher in the southern and the western universities, respectively. For courses approaching the largest in the sample (i.e., 200 students in the southern, 400 in the western) a significant gap emerges, with male instructors scoring half a point higher. Given differences in course sizes, average evaluations, and wording of questions, the estimated effect of interaction between gender and course size is remarkably consistent across the two universities. Differences of this magnitude are large enough to capture the attention of promotion and tenure committees, award committees, and the like. For universities that offer even larger classes, the cumulative effect would be massive. Although this particular study is based on only two universities, it is consistent with studies in other fields and with the theoretical literature on role incongruity. It shows a systematic and sizeable bias against female instructors in large courses.

Given differences in course sizes, average evaluations, and wording of questions, the estimated effect of interaction between gender and course size is remarkably consistent across the two universities. Differences of this magnitude are large enough to capture the attention of promotion and tenure committees, award committees, and the like.

What difference does this apparent bias make? Of course, it depends on institutional practice. The worst-case scenario includes exclusive or predominant reliance on SETs for assessment of

Table 1

Effect of Course Size and Gender of Instructor on Average Course Evaluation

	Coefficient	Standard Error	t-Statistic	95% Confidence Interval	
Southern University					
Intercept**	4.48	0.0962	46.55		
Number of respondents**	-0.00501	0.00184	-2.72	-0.000862	-0.00139
Number x Male*	0.00372	0.00192	1.94	-0.0000527	0.00749
Male instructor	-0.199	0.106	-0.190	-0.228	0.188
N = 309					
Western University					
Intercept**	4.20	0.0465	90.29		
Enrollment**	-0.00151	0.000356	-4.24	-0.00221	-0.000812
Enrollment x Male**	0.000974	0.000377	2.58	0.000234	0.00171
Male instructor	-0.0133	0.0533	-0.25	-0.118	0.0914
N = 587					
Tobit analysis	* = significant at 0.10 level	** = significant at 0.05 level			

Table 2

Estimated Average Teacher Effectiveness Score, Five-Point Scale

Southern University	Course Size = 10	Course Size = 100	Course Size = 200
Female instructor	4.43	3.98	3.46
Male instructor	4.27	4.15	4.01
Western University	**Course Size = 10**	**Course Size = 100**	**Course Size = 400**
Female instructor	4.18	4.05	3.58
Male instructor	4.18	4.13	3.95

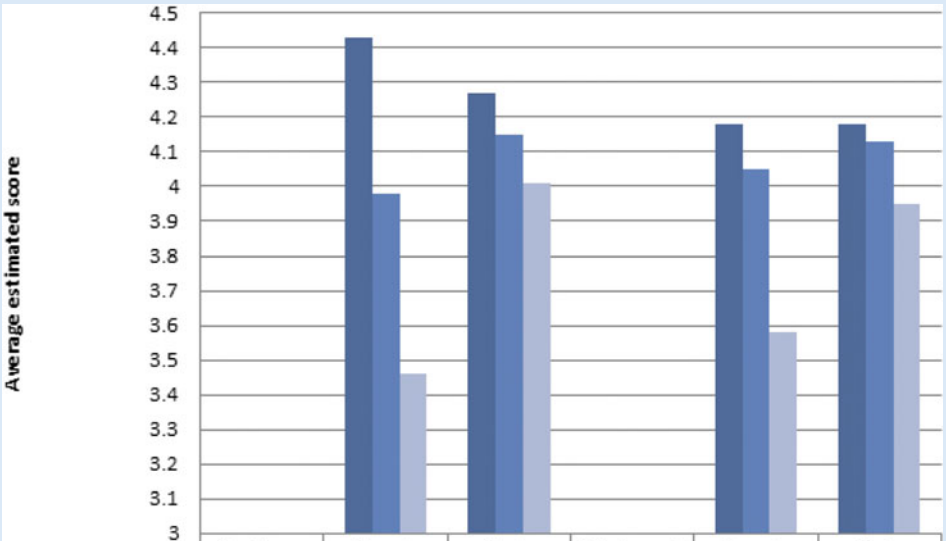
teaching effectiveness; emphasis on success in teaching larger courses; and a prominent role for teaching evaluations in professional advancement. Whereas these conditions do not hold in all or perhaps even most political science departments, they are not uncommon. For example, decisions about retention of adjunct faculty often are based solely on SETs; therefore, individual careers are wholly dependent on this one apparently biased measure.

An immediate effect of bias is likely that women disproportionately teach smaller courses than men. This could result from several mechanisms: women self-selecting out of teaching large courses; departments channeling women into teaching smaller courses; and students selecting into lectures that are taught by men. I do not take a stance on what the causal mechanism is; however, to the extent that successful teaching of large classes provides material or other rewards within departments, any process that leaves women disproportionately teaching small classes is an impediment to professional advancement. In the datasets analyzed here, there is evidence of women systematically teaching smaller courses than men. The mean course size for female faculty at the southern university is 34 students; for male faculty, it is 51 students. In the western university, courses taught by female faculty have an average size of 91 students; in those taught by male faculty, it is 123 students. A two-sample t-test shows that these are statistically significant differences in mean course size.[2]

Given increasing evidence on gender bias in SETs, it is time for the pendulum to swing in the other direction: away from telling women to lean in and to perform better within the current system and toward developing better metrics of teaching effectiveness.

Figure 1

Average Evaluation Score



More than 30 years ago, Martin (1984) wrote that the "message to women faculty seems clear: if your institution bases personnel decisions on student evaluations, make sure your colleagues are aware of the possibility of sex bias" (Martin 1984, 492). Three decades later, we essentially use the same evaluation tools, and colleagues remain skeptical of the presence of gender bias. Specifically for evaluations of women faculty in large courses, bolstered by studies in other disciplines, we find that the bias is strong and must be considered by departments and universities.

CONCLUSION

Recent public debate about women's professional advancement

has fallen into a dichotomy between those who argue that ambitious women need to "lean in" and those who draw attention to structural and implicit biases that work against women's success at the highest levels. This current debate has direct relevance to the topic of this article. Gender interacts with aspects of the classroom environment to influence SETs. In particular, when women assume a stereotypical leadership role, as in a large lecture course, beliefs about gender and leadership have an impact on evaluations of teaching effectiveness. The evidence presented in this article supports this hypothesis and questions the use of SETs in consideration of promotion, compensation, awards, prominent administrative positions, and similar tokens of professional success. As Boring (2015, 6–7) concluded: "[S]tudents are not evaluating teachers' helpfulness in making them learn when they complete their evaluations.... And yet, universities continue to use this tool in a way that may hurt women (and probably other minorities as well, and men who do not correspond to students' expectations in terms of gender stereotypes) in their academic careers."

Regarding the lean-in versus structural impediments dichotomy, the literature so far has fallen heavily on the former. Publications in political science journals (as well as in other disciplines) offer advice on how female faculty can increase their scores on SETs. Women have reported engaging in tactics to show their sensitivity to student needs and to illustrate their "niceness." Many also take steps to better project their authority and competence, such as by participating in acting workshops. They spend considerable time on course preparation and organization. Some of these steps increase actual teaching effectiveness. However, faculty members—male and female—acknowledge that SETs can be gamed, and they offer advice on how to do so. Therefore, we are all encouraged to take the existing evaluation system as given and to lean in.

Given increasing evidence on gender bias in SETs, it is time for the pendulum to swing in the other direction: away from telling women to lean in and to perform better within the current system and toward developing better metrics of teaching effectiveness. For example, when we consider teaching effectiveness for graduate courses, we might consider SETs. However, a far more persuasive and widely used indicator of whether a professor is effective in training graduate students is results: Do the professor's students obtain good jobs and go on to become prominent figures in the profession? To the extent that we can move away from SETs as a sole or primary indicator of teaching effectiveness at the undergraduate level and emulate what we naturally do at the graduate level, our assessments would be more reliable. Some institutions have moved toward a process of peer review to complement SETs. Although this innovation makes some faculty uncomfortable, peer review by faculty members who are given advice on how to do it well could be a substantial improvement on the currently dominant system (Stark and Freishtat 2014). Evaluation by trained observers is another possibility, although it would require investment by universities.

It also is possible that in some settings, more objective measures of teaching success could be developed. If multiple sections of the same course are taught by different faculty, for example, it may be possible to ask students to engage in a form of standardized assessment of how much they have learned. Effectiveness in teaching large introductory courses could be measured by assessing how well students perform later in more advanced courses. One recent study examined such a setting, in which economics students at Bocconi University were randomly assigned in introductory economics courses (Braga, Paccagnella, and Pellizzari 2014). The authors found that, indeed, SETs are significantly correlated with success in more advanced courses—*but in the wrong direction*. That is, teachers who receive lower ratings produce students who go on to achieve higher grades in advanced classes.[3]

Of course, none of these changes could be implemented immediately or without controversy. However, given the long-standing concerns about heavy reliance on SETs, theory that bolsters these concerns, and evidence of bias in SETs in political science, change is long overdue. Questions about how new assessment technologies might work is no excuse for continuing to rely on existing mechanisms that are known to be faulty. We have enough advice on how to lean in; it is time to make structural changes.

ACKNOWLEDGMENTS

The work presented in this article was influenced by many conversations with colleagues. I particularly thank Eve Fine, Yoi Herrera, Bob Keohane, Helen Kinsella, Rose McDermott, Ryan Powers, and Barbara Walter. Any mistakes, of course, are solely my responsibility. ■

NOTES

1. Using the number of responses as an estimate of enrollment could raise concerns of bias in the results if response rates are systematically correlated with other variables of interest. However, an internal investigation by this university mitigates this concern. It found that the response rate does not affect the mean instructor rating, the variable being measured here. The reason is that response bias is likely to occur at both ends of the scale—students who strongly liked and strongly disliked the course are more likely to respond. Thus, the mean score is not likely to be influenced by the response rate.
2. The difference in means in the southern university has a t-value of -2.05 and in the western university of -3.53. Both are significant at the 0.05 level.
3. The authors also found, disturbingly, that SETs are significantly correlated with "meteorological conditions."

REFERENCES

Andersen, Kristi, and Elizabeth D. Miller. 1997. "Gender and Student Evaluations of Teaching." *PS: Political Science & Politics* 30: 216–18.

Arbuckle, Julianne, and Benne D. Williams. 2003. "Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations." *Sex Roles* 49: 507–16.

Baldwin, Tamara, and Nancy Blattner. 2003. "Guarding Against Potential Bias in Student Evaluations: What Every Faculty Member Needs to Know." *College Teaching* 51: 27–32.

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: 308–14.

Bianchini, Stefano, Francesco Lissoni, and Michele Pezzoni. 2012. "Instructor Characteristics and Students' Evaluations of Teaching Effectiveness." *European Journal of Engineering Education*, iFirst: 1–20.

Boring, Anne. 2015. "Gender Biases in Student Evaluations of Teachers." Working paper, OFCE-PRESAGE-Sciences Po and LEDa-DIAL (France).

Braga, Michele, Marco Paccagnella, and Michele Pellizzari. 2014. "Evaluating Students' Evaluations of Professors." *Economics of Education Review* 41: 71–88.

Butler, Dore, and Florence L. Geis. 1990. "Nonverbal Affect Responses to Male and Female Leaders: Implications for Leadership Evaluations." *Journal of Personality and Social Psychology* 58: 48–59.

Centra, John A., and Noreen B. Gaubatz. 2000. "Is There Gender Bias in Student Evaluations of Teaching?" *The Journal of Higher Education* 70: 17–33.

Dion, Michelle. 2008. "All-Knowing or All-Nurturing? Student Expectations, Gender Roles, and Practical Suggestions for Women in the Classroom." *PS: Political Science & Politics* 41 (4): 853–6.

Eagly, Alice H., and Steven J. Karau. 2002. "Role Congruity Theory of Prejudice toward Female Leaders." *Psychological Review* 109: 573–98.

Feldman, Kenneth A. 1993. "College Students' Views of Male and Female College Teachers: Part II—Evidence from Students' Evaluations of Their Classroom Teachers." *Research in Higher Education* 34: 151–211.

Johnson, Michael D., Arunachalam Narayanan, and William J. Sawaya. 2013. "Effects of Course and Instructor Characteristics on Student Evaluation of Teaching across a College of Engineering." *Journal of Engineering Education* 102 (2): 289–318.

Johnson, Stefanie, Susan Murphy, Selamawit Zewdie, and Rebecca 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: 39–60.

Langbein, Laura I. 1994. "The Validity of Student Evaluations of Teaching." *PS: Political Science & Politics* 27 (3): 545–53.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2014. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40 (4): 291–303.

Martin, Elaine. 1984. "Power and Authority in the Classroom: Sexist Stereotypes in Teaching Evaluations." *Signs* 9: 482–92.

Miller, Joann, and Marilyn Chamberlin. 2000. "Women are Teachers, Men are Professors: A Study of Student Perceptions." *Teaching Sociology* 28 (4): 283–98.

Ridgeway, Cecelia L. 2001. "Gender, Status, and Leadership." *Journal of Social Issues* 57: 637–55.

Rudman, Laurie A., and Peter Glick. 2001. "Prescriptive Gender Stereotypes and Backlash toward Agentic Women." *Journal of Social Issues* 57: 743–62.

Sampaio, Anna. 2006. "Women of Color Teaching Political Science: Examining the Intersections of Race, Gender, and Course Material in the Classroom." *PS: Political Science & Politics* 39 (4): 917–22.

Sidanius, Jim, and Marie Crane. 1989. "Job Evaluation and Gender: The Case of University Faculty." *Journal of Applied Social Psychology* 19: 174–97.

Smith, Holly. 2012. "The Unintended Consequences of Grading Teaching." *Teaching in Higher Education* 17: 747–54.

Stark, Philip B., and Richard Freishtat. 2014. "An Evaluation of Course Evaluations." Manuscript, University of California, Berkeley.

Wigington, Henry, Nona Tollefson, and Edme Rodriguez. 1989. "Students' Ratings of Instructors Revisited: Interactions among Class and Instructor Variables." *Research in Higher Education* 30: 331–44.