Gender, Teaching Evaluations, and Professional Success in Political Science

Lisa L. Martin

University of Wisconsin-Madison

llmartin3@wisc.edu

Prepared for the American Political Science Annual Meeting, August 29-September 1, 2013, Chicago, IL.

The work presented in this paper was influenced by many conversations with colleagues. I would particularly like to thank Yoi Herrera, Bob Keohane, Barbara Walter, Rose McDermott, and Eve Fine. Any mistakes are of course solely my responsibility, as are any opinions rashly expressed.

As part of this theme panel on "Minding the Gender Gap in Political Science," I hope to direct our attention to the role of teaching and evaluations of teaching in professional advancement in the discipline. Many female faculty believe that they face prejudice in student evaluations of teaching (SETs), and that this prejudice may be exaggerated by developments such as online evaluations and the prevalence of sites such as RateMyProfessor. However, systematic studies of SETs are mixed in their findings of gender bias.

In this paper I make a number of interrelated arguments. First, a review of the psychological literature on gender and leadership assessments suggests that there is an interaction between gender 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 or a Massive Open Online Course (MOOC) – assumptions about gender roles could have a significant impact on evaluations. Second, I provide a preliminary assessment of this hypothesis using publicly-available SET data from a political science department at a large public university. These data suggest, as expected, that female faculty receive lower evaluations of general teaching effectiveness in large courses than do male faculty, while no substantial difference exists 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 may be biased creates concerns. Third, the race by universities to join the MOOC game so far has exhibited a strong preference for courses taught by male faculty. All of these concerns suggest that the discipline needs to reconsider its methods of faculty evaluation and the role that such evaluations play in professional advancement.

Role Congruity, Leadership, and Evaluation of Teaching Effectiveness

The potential for gender bias in SETs has long been recognized and discussed. One seminal study (Ambady and Rosenthal 1993) that has gained much attention highlights the question of what teacher effectiveness ratings are really capturing. The authors are interested in the effect of "thin slices" of nonverbal behavior on observers' global evaluations. Their primary study showed student judges 30-second silent video clips of college teachers at a private university. They asked students to rate these teachers along 15 dimensions of nonverbal behavior, including terms such as attractive, competent, enthusiastic, and supportive. They then correlated these judgments with the end-of-semester overall ratings of effectiveness for these teachers. The authors found a statistically significant positive correlation with nearly all 15 dimensions, and together these dimensions explained 76% of the variation in overall evaluations. Further studies exposed judges to even shorter video clips – 6 and 15 seconds – and likewise found significant correlations. If a 6-second silent clip explains most of the variation in overall teacher effectiveness ratings, what are these ratings really capturing? In what sense are they valid indications of teaching effectiveness?

Within political science, the APSA has occasionally published a piece in *PS* that draws attention to the potential for bias in SETs, and offers advice for faculty who are worried about it. 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 questions whether SETs are adequately valid measures of teaching effectiveness to play such an important role. Andersen and Miller (1997) note that female instructors who are not perceived as caring and accessible may fail to meet student expectations, and so be penalized on SETs. Sampaio (2006) examines the intersection of

gender, race, and subject matter, focusing on implications for women of color in the classroom. Dion (2008) reviews the literature on bias and offers advice for women faculty who must be both authoritative and nurturing. She suggests that, given reliance on possibly biased SETs, women faculty need to focus on course management, making a good impression on the first day of class, explaining grading rubrics in exhaustive detail, and expressing interest and concern for students as well as exhibiting a friendly demeanor. In related work, Baldwin and Blattner (2003) suggest that since SETs may be biased, alternative evaluation measures should be considered. Smith (2012) notes that SETs are used both for professional development and for employment decisions, setting up tensions. These tensions are especially pronounced given questions about the validity and reliability of SETs, as well as of peer observation of teaching.

The evidence on gender bias in SETs more generally, however, is mixed. This is perhaps one reason that concerns about reliance on SETs have not had more of an impact on university practice. Sidanius and Crane (1989) did find an across-the-board effect by which women faculty received lower ratings. In looking at ratings of over 400 faculty by more than 9000 students, they came to three major conclusions. First, males scored significantly higher on evaluations of global teacher effectiveness. Second, after controlling for other variables, they found that women are not perceived by students as more sensitive to students' needs (which students could potentially hold against these women, if they anticipate that women will be more sensitive). Third, in looking at the determinants of overall effectiveness ratings, Sidanius and Crane find that academic competence weighs more heavily on ratings for male than for female faculty.

These findings echoed those of Basow and Silborg (1987). Basow (1995) is able to look at the effect of student gender on teacher ratings. She finds that ratings of male faculty are not influenced by student gender. However, female faculty are rated lower by male students than by

female students. Centra and Gaubatz (2000), in a related study, find small same-gender preferences but not globally lower ratings for women instructors. Kierstad, D'Agostino, and Dill (1988) examine some of the factors that go into overall effectiveness evaluations. Like most other studies of the time, they find that male faculty are overall rated more effective. They also find that perceptions of friendliness enter into these ratings, as being perceived as friendly improves women faculty's ratings, but has no effect on men's. Feldman (1993) casts doubt on studies that find significantly lower ratings for female than male faculty, arguing that overall, studies find small and inconsistent differences. Using data from RateMyProfessor.com, Reid (2010) likewise finds small and inconsistent gender effects. He finds that race matters much more, with black male faculty receiving the most negative ratings. In a recent study undertaken in an Italian engineering college, Bianchini, Lissoni, and Pezzoni (2012) do find significant gender effects. In three out of the four programs that they examine, women consistently receive significantly lower effectiveness scores than men. The authors speculate that the gender makeup of the student body could account for their findings, as two of the four programs have very low percentages of female students.

In the psychology literature, some interesting experiments shed light on these observational studies. Freeman (1994) distinguishes between stereotypically male (authoritative, competent) and female (nurturing, approachable) gender roles. In an experimental approach, he finds that female instructors only received high ratings if they scored highly on both gender roles. Men only needed to get a high score on the male gender role to receive a high rating. Arbuckle and Williams (2003) undertook a fascinating experiment in which students viewed a stick figure that delivered a short lecture. While all participants observed the same stick figure and same lecture, 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, illustrating that student expectations influence their perception of an instructor independent of the material being delivered, or how it is delivered.

Turning to the more general literature on gender and leadership, a body of work known as "role congruity theory" helps to put these studies of SETs in context and to suggest 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, being more passive, nurturing, and sensitive. Role incongruity occurs when a man or woman acts in a way that is contrary to type, for example if a women takes on an agentic demeanor. A situation that demands that a woman be agentic, for example teaching a large lecture class (as opposed to a small seminar) will cause role incongruity and can lead to negative reactions from students.

Butler and Geis (1990) use experimental approaches to examine the role of gender and leadership in the reactions of observers. They focus on nonverbal responses, in particular positive or negative facial reactions of participants who observe leaders making suggestions for certain courses of action. Male and female leaders made the same suggestions. However, the female leaders elicited significantly more negative facial expressions than did men in the same situation. Ridgeway (2001) discusses "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. Such beliefs may shape the likelihood that women emerge as leaders in the first place. Experiments also reveal that when women are

placed in a leadership role and act assertively, they are punished. Along similar lines, Rudman and Glick (2001) examine the potential for backlash against agentic women. In their work, they find 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" (743).

Eagly and Karau (2002) review the work on role congruity theory and female leadership. They conclude that two forms of prejudice are most prominent. First, women are in general viewed less favorably as leaders. Second, when women exhibit behaviors that are generally associated with leadership, such as projecting authority, women are evaluated less favorably than men. Johnson et al. (2008) conduct a series of tests of role congruity theory, using qualitative, experimental, and survey approaches. They contrast the "strong" (agentic) type to the "sensitive" (non-agentic) type. Consistent with other studies, they find that female leaders need to project both strength and sensitivity to be effective, while male leaders need only to project strength.

Taken as a whole, these studies suggest a more nuanced approach to the potential for gender bias in SETs. The effect of gender on overall ratings may be negative, but is not large and studies are not consistent. However, different kinds of courses demand that instructors take on different roles. In small classes, such as seminars, the instructor is usually seated and his or her role is to guide discussion and draw out students' thoughts, as well as facilitating class discussion. In this setting, students likely do not come to class with expectations that the instructor play the typical agentic leader role. However, contrast a large lecture course, with the instructor on a stage with a microphone in front of hundreds of students. The opportunities for interaction with individual students are limited, as are chances to express concern for their

individual needs or to draw out their opinions. Instead, students are likely to come into class with standard expectations of agentic leadership.

If this is the case, the potential for backlash against agentic women will be high in these large lecture settings, while it is likely to be minimal or absent in small class settings. This hypothesis could explain the inconsistent results on gender bias in SETs. Perhaps these biases only 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 all this is correct, what we need to look for in order to test the hypothesis is an interaction effect between class size and lower effectiveness ratings for female faculty. The presence of such an effect would validate role congruity theory's relevance to the classroom, and renew concerns about reliance on SETs as measures of teaching effectiveness. The next section presents some evidence testing this hypothesis.

Some Evidence and Implications

The implications of studies of gender roles and leadership for evaluation of teaching effectiveness seem clear: when the teaching environment places women in a stereotypical leadership role, role incongruity will lead to lower evaluations of effectiveness. One observable implication follows directly. In a large lecture course, where the format is the "sage on the stage" who needs to project authority and where personal interaction with students is minimal, students are likely to rate female teachers as less effective than male teachers. We would not expect, based on leadership studies, that such an effect would obtain in a smaller, more

facilitative setting such as a seminar. Thus, a direct implication is that we should find an interaction effect between gender and the size of courses.

While a number of types of interaction effects between gender and other characteristics of courses have received attention, this specific interaction between course size and instructor gender has not been studied in any depth. One exception is Wigington, Tollefson, and Rodriguez (1989), which collected data involving 5843 student evaluations at a Midwestern university in the mid-1980s. These 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…" (339). This effect reversed for small classes. Unfortunately, the authors did not pursue this result any further and it seems to have gotten lost in the general sense that "interactions matter."

Updating the Wigington, Tollefson, and Rodriguez study is difficult because nearly all universities now make their SETs available only to faculty within the university (and sometimes to students). I have analyzed the data from the last ten years in my own department, and have found a large and statistically significant interaction effect consistent with their study. (Any interested reader could presumably conduct the same analysis within their own department.) However, this result is based on data that are not publicly available and so cannot be replicated; and obviously they reflect only one department's experience.

Recent trends in attempting to hold public universities more "accountable" have had the possibly unintended effect of making more teaching evaluation data publicly available. Without endorsing these accountability efforts, I am able to take advantage of new data availability. The following analysis is based on records from a political science department in a large, public

southern university for Fall of 2011 through Spring of 2012.[1] Table 1 examines the effect of gender, course size, and the interaction between the two on average course evaluations.

[Table 1 about here]

The dependent variable in this analysis is the average response, on a 5-point scale, to the statement: "Overall this instructor was effective." "Strongly agree" is equivalent to 5 points, and "strongly disagree" is equivalent to 1 point. Analysis is based on all 198 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.

The coefficients go in the expected direction, with in particular a positive interaction effect between having a male instructor and a larger class. Given the sample size, and especially the relatively small number of courses taught by female instructors in this sample, the coefficients unfortunately do not meet tests of statistical significance; in further work, I hope to increase the sample size by finding other publicly-available data sources. However, the substantive effects are interesting, in the expected direction, and suggest that further research on this question is warranted. Table 2 summarizes the estimated substantive effects.

[Table 2 about here]

For a small course, with ten students, there is only a tenth of a point difference in ratings between male and female instructors. For a larger course of 100 students, a more sizeable difference emerges, with males now scoring three-tenths of a point higher. For a course of 200 students – near the largest size in this sample – a large gap emerges, with male instructors scoring six-tenths of a point higher. Differences like this are large enough to catch the attention of promotion and tenure committees, award committees, and the like. For universities that offer

[1] I welcome suggestions of other departments whose data are publicly accessible so that I can increase the sample size.

even larger classes – say, 400 or 900 students (think of Michael Sandel's famous Justice course at Harvard, now a MOOC) – the cumulative effect would be massive. While this particular study is far from definitive, it is consistent with earlier studies and with the theoretical literature on role incongruity. It suggests a systematic and sizeable bias against female instructors in large courses.

What difference does this apparent bias make? It of course depends on institutional practice. The worst-case scenario would include exclusive or predominant reliance on SETs for assessment of teaching effectiveness; emphasis on success in teaching larger courses; and a prominent role for teaching evaluations in processes of professional advancement. While these conditions do not hold in all, or perhaps even most, political science departments, they are probably not uncommon.

One immediate effect of bias would likely be that women disproportionately teach smaller courses than men. This could result from a number of mechanisms: women self-selecting out of teaching large courses; departments channeling women into teaching smaller courses; or students selecting into lectures that are taught by men. I do not have a stand on what the causal mechanism is, but 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 becomes an impediment to professional advancement. In the dataset analyzed here, we do see evidence of women systematically teaching smaller courses than men. The mean course size for female faculty is 31.6 students; for male faculty, 49.7 students (remembering that course size is estimated by the number of respondents in this sample).

As an initial attempt to investigate the method and role of teaching evaluations in political science departments, I conducted a small online survey. I used a snowball survey

method, running the survey in June and July 2013. The survey resulted in 117 responses. This survey methodology does not produce a random sample of political scientists, but does provide some insight and a foundation for further analysis.

First, I asked about what methods are used to evaluate faculty teaching. 69% of respondents reported that their department used student evaluations completed online; 50% SETs completed in class; and 32% peer evaluations. Only 3% reported any use of objective measures such as standardized tests. In comments, a number of respondents reported that even though the department conducted peer evaluations or looked at a teaching portfolio, tenure evaluations focused solely on SETs.

Next, I asked about the attention that teaching evaluations receive in department discussions and the role that they play in promotion reviews. Only 39% of respondents said that teaching evaluations received "frequent" or "moderate" attention in faculty discussion, but nearly 78% said that evaluations were "very important" or "somewhat important" in promotion reviews. One specific concern expressed in comments by a number of respondents was the role of SETs in nominations for university teaching awards. SETs often play a dominant role in the nomination process (one that carries with it monetary and prestige awards) to the exclusion of other indicators of teaching excellence such as innovation or mentorship. Comments included: teaching evaluations are "heavily used in determining teaching awards, which women rarely receive"; "used when making departmental nominations for university-wide awards (e.g., term professorships) or departmental awards…."

Many respondents added comments about the importance of teaching evaluations in tenure decisions, and these comments indicate the tremendous variety of institutional practice in this crucial decision. At one extreme, some say that evaluations are always considered but never

decisive; at the other, respondents indicating that they were at liberal arts colleges or professional schools indicated that teaching evaluations could sink a tenure case or put it over the line. A number of respondents indicate that teaching evaluations come into play at the extremes, when they are exceptionally good or exceptionally bad. One potentially troubling type of comment made by a large number of respondents regards de facto exclusive use of SETs as indicators of teaching excellence during tenure decisions: "Generally evaluations are the only criteria used to evaluate untenured faculty's teaching abilities"; "Teaching evaluations are the only criteria used to determine how faculty are performing in the classroom. High averages earn faculty high rankings, which in turn results in high faculty evaluations and the possibility for promotion and raises"; "Teaching evaluations are the only tool used to assess the teaching component of the tenure and promotion file…."

I also asked about the importance of teaching evaluations for reward systems other than tenure, such as awards, raises, and selection to leadership positions. Here, 16.5% felt that evaluations were "very important" in this context and 47.4% that they were "somewhat important." Again, written comments revealed tremendously different practices, some saying that evaluations had "no relationship" to such decisions and that evaluations "are, basically, irrelevant after someone gets tenure," but another that "40% of annual raises come down (effectively, at least) to numerical teaching evaluations administered online."

I also asked questions about perceived gender and other types of bias in teaching evaluations, recognizing of course that this survey could only elicit subjective impressions. On the question of gender bias, the most common response was "not sure" (45%). The survey did not mention course size as a factor in evaluations, but nonetheless respondents frequently noted a perceived relationship between leadership expectations and gender. "I personally experience

worse evaluations in large classes than my male peers who teach the same course"; "a colleague and I teach the same large required intro course, we are about the same age/rank, we have a very similar approach (requirements, readings, approach to the material), and a similar grade distribution. But the perception among students (and this comes through in evaluations too; and my colleague has won numerous teaching awards) is that I'm much tougher/meaner, and that my class should be avoided in favor of his"; "Aside from one woman who 'entertains' the students in intro classes, women consistently receive lower evaluations. This is especially true for those who are perceived as 'tough'"; "I have noticed that students often personalize aspects of the student-professor relationship with a female professor in ways that they do not with male professors'"; "I'm a tall male, I get no guff. But my colleagues, women who have a commanding presence generally, get far more challenge to their authority."

The literature on role congruity and leadership leads us to expect that when female faculty are placed in a stereotypical situation of needing to project authority – such as lecturing to a large class – students will find them less effective than male faculty in the same situation. The evidence presented here is based on a relatively small sample, but shows a large substantive effect in the predicted direction. Survey responses also indicate that these SETs play an important role in department and university decisionmaking, at least at some institutions. This confluence of factors suggests that the evaluation tools being used in crucial promotion and award decisions could be systematically biased against female faculty. Nearly thirty years ago, Elaine 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" (492.) Thirty years on, we largely use the same evaluation tools, and

colleagues remain skeptical of the presence of sex bias. Looking specifically at evaluations of women faculty in large courses, increasing evidence of such bias is emerging.

MOOCs: Intensifying the Dilemma

As we all know, higher education is facing fundamental challenges with the advent of online and distance learning. Even with grave concerns about MOOCs, universities are racing to offer them, joining consortia such as Coursera, Udacity, and edX. The potential impact of a rush to MOOCs, and online learning more generally, on gender disparities has not yet received much attention. However, the line of analysis that I've spelled out so far in this paper suggests possibly major effects.

Theory and evidence suggest that the "sage on a stage" model of education has negative consequences for female faculty, as it exaggerates and reinforces role incongruity. The leadership and authority that faculty must project in such a situation clash with expectations of women being more nurturing. As we move to online technologies, lecturing to ever larger audiences via distance learning and MOOCs, we should only expect these biases and effects to be exaggerated. Course sizes can become enormous, and individual interaction between instructors and students during lectures is eliminated. In a peculiar way, the movement to MOOCs reinforces a mode of learning that otherwise was coming to seem dated, with one authoritative figure lecturing to large groups of passive learners. As one journalist who took eleven MOOCs and reported on them for the *New York Times* says, the "professor is, in most cases, out of students' reach, only slightly more accessible than the pope or Thomas Pynchon" (Jacobs 2013).

From the perspective of the role of women faculty in universities, to the extent that universities offer MOOCs, and that those who teach MOOCs receive rewards and acknowledgment within the structure of the university, the dilemmas and biases identified in this paper become more intense. The impersonal, remote structure of a MOOC seems likely to reinforce students' gender stereotypes, and to make it even more difficult for female faculty teaching such courses to walk the fine line between projecting authority and being friendly.

If this analysis is correct, we might expect that the early jumps into MOOCs have tended to feature predominantly male faculty. A number of causal mechanisms could be at work. One might be that when university approach faculty, asking them to offer MOOCs, they naturally go initially to professors who have received high ratings in large lecture courses. Another could be that women who have found teaching such large courses challenging will self-select out of offering MOOCs. It is also possible that student expectations could mean that they tend to gravitate toward MOOCs that are offered by men. Perhaps all of these mechanisms, and more, are at work. Regardless of the specifics of the process, if men overwhelmingly teach MOOCs, increasing reliance on this form of instruction is likely to be detrimental to the professional advancement of female faculty.

The evidence on MOOCs so far shows that, as expected, they are largely taught by men. In an op-ed piece we published in the *Los Angeles Times* in January 2013, Barbara Walter and myself presented some of the numbers (Martin and Walter 2013). Coursera, one of the largest of the MOOC providers, at that time offered 205 courses. Of these, only 35 were taught by female faculty. 157 were taught by male faculty, and the remaining 14 by mixed gender groups. Udacity, another large MOOC provider, offered almost no MOOCs with any female instructors. Looking in more depth at some of the universities that participate in Coursera, even larger gender

disparities emerge. Princeton University has 33% female faculty; however, none of them participate in the MOOCs Princeton offers through Coursera. The University of Pennsylvania, also with 33% female faculty, has only 12.5% of its MOOCs being taught by women. The efforts that prominent universities have made in the last decades to hire and promote more women faculty are so far not reflected in their movement toward MOOCs.

A relatively new entry in the MOOC provider game is edX, with leading roles being played by Harvard University and MIT. Unfortunately, the story of edX is the same as for other MOOC providers, perhaps even more extreme, in spite of edX's commitment "to deliver these teachings from a faculty who reflect the diversity of its audience" (edX mission statement). Of the initial 25 courses that edX listed on its website, none are taught solely by female faculty. 17 are taught solely by male faculty, and the remaining 8 by mixed groups.[2] These gender disparities cannot be explained by the fact that many MOOCs are in subjects such as mathematics where faculty members are disproportionately male. Even in fields with higher percentages of female faculty, such as in the humanities, nearly all MOOCs are taught by men.

Walter and Martin (2013) focus on the implications of the missing female MOOC instructors for developing countries. Volumes of research have demonstrated that the education of women and girls in developing countries is crucial to processes of social and economic development, and developing countries such as India are voracious consumers of MOOCs. By not putting women at the head of these courses, we are missing an opportunity to export valuable female role models of authority and leadership.

However, for purposes of this paper I would emphasize the impact of the MOOC dynamic on the professional advancement of women faculty. MOOCs provide prestige and

[2] "edX and Gender Equity," *Harvard Magazine*, July-August 2013, p. 8.

visibility within the university and discipline. As A.J. Jacobs reports, "MOOCs are creating a breed of A-list celebrity professors who have lopsided sway over the landscape of ideas" (Jacobs 2013). And these "celebrity professors" are overwhelmingly male. One expects that one day, if not already, they will also provide financial rewards to faculty.[3] To the extent that women faculty have limited access to these rewards, the movement toward MOOCs undermines efforts to advance women within the academy. Large courses within the university already present major challenges for female faculty; these challenges seem exaggerated as we move rapidly toward online educational models.

Conclusion

The 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, like Anne-Marie Slaughter, who draw attention to structural and implicit biases that work against women's success at the highest levels. Like many, I find this dichotomy usually strained and not helpful. However, it does have direct relevance to the topic of this paper: the role of teaching effectiveness and student evaluations of teaching in the advancement of female faculty. 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 can be expected to have an impact on evaluations of teaching effectiveness. While preliminary, the evidence presented in this paper supports this hypothesis and calls into

[3] Consider, as a relevant precedent, the *Great Courses* series of dvds that is promoted heavily in venues such as *The Economist* magazine. The *Great Courses* website currently lists 249 professors; of these, 29, or just 11.6%, are women.

question the use of SETs in consideration of promotion, compensation, awards, selection into prominent administrative positions, and similar tokens of professional success.

What to do? Returning to the lean-in vs. structural impediments dichotomy, the literature so far has fallen heavily on the lean-in side. Publications in political science journals (as well as in other disciplines) offer advice on how female faculty can increase their scores on SETs. Women report engaging in tactics to show their sensitivity to student needs and to illustrate their "niceness," such as inviting entire classes over for Thanksgiving dinner. Many also take steps to better project their authority and competence, such as participating in acting workshops. Some of these steps surely do increase actual teaching effectiveness. However, faculty – male and female – acknowledge that SETs can be gamed, and offer advice on how to do so. So we are all encouraged to take the existing evaluation system as given, and to lean in.

Given the persistent reliance on 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. Some institutions have moved toward a process of peer review to complement SETs. While this innovation makes some faculty uncomfortable, peer review by faculty who are given advice and training on how to do it well could be a substantial improvement on the currently dominant system. It is also 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 some form of standardized assessment of how much they have learned. Effectiveness in teaching large introductory courses could perhaps be measured by assessing how well students perform later in more advanced courses. Of course, none of these changes could be implemented immediately or without controversy. However, given longstanding

concerns about heavy reliance on SETs, theory that bolsters these concerns, and emerging evidence of bias in SETs in political science, it is past time for a change. Questions about how new assessment technologies might work is no excuse for continuing to rely on existing mechanisms that we know are faulty. Enough advice on how to lean in, time to make some structural changes.

Table 1

Effect of Course Size and Gender on Average Course Evaluation

	Coefficient	Std. error	t-statistic
Intercept**	4.36	0.124	35.3
Number of respondents	-0.00372	0.00249	-1.50
Number x Male	0.00270	0.00259	1.04
Male instructor	0.0842	0.135	0.623
N=198			
r-square=0.033			

Table 2

Estimated Average Teacher Effectiveness Score, 5-point scale

	Course size = 10	Course size = 100	Course size = 200
Female instructor	4.33	3.99	3.62
Male instructor	4.43	4.34	4.24

References

Ambady, Nalini, and Robert Rosenthal. 1993. "Half a Minute: Predicting Teacher Evaluations from Thin Slices of Nonverbal Behavior." *Journal of Personality and Social Psychology* 64: 431-44.

Andersen, Kristi, and Elizabeth D. Miller. 1997. "Gender and Student Evaluations of Teaching." *PS: Political Science and 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. 1995. "Student Evaluations of College Professors: When Gender Matters." *Journal of Educational Psychology* 87: 656-65.

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.

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 and Politics* 41: 853-56.

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.

Freeman, Harvey R. 1994. "Student Evaluations of College Instructors: Effects of the Type of Course Taught, Instructor Gender and Gender Roles, and Student Gender." *Journal of Educational Psychology* 86: 627-30.

Jacobs, A.J. 2013. "Two Cheers for Web U!" *New York Times,* April 20: SR1.

Johnson, Stefanie, et al. 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.

Kierstead, Diane, Patti D'Agostino, and Heidi Dill. 1988. "Sex Role Stereotyping of College Professors: Bias in Students' Ratings of Instructors." *Journal of Educational Psychology* 80: 342-44.

Langbein, Laura I. 1994. "The Validity of Student Evaluations of Teaching." *PS: Political Science and Politics* 27: 545-53.

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

Martin, Lisa L., and Barbara F. Walter. 2013. “Setting an Online Example in Educating Women.” *Los Angeles Times,* January 25.

Reid, Landon D. 2010. “The Role of Perceived Race and Gender in the Evaluation of College Teaching on RateMyProfessor.com.” *Journal of Diversity in Higher Education* 3:137-52.

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 and Politics* 39: 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.

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.