


RESEARCH

Open Access



Gender patterns in engineering PhD teaching assistant evaluations corroborate role congruity theory

C. A. Evans^{1*} , K. Adler², D. Yucalan³ and L. M. Schneider-Bentley¹

Abstract

Background The body of work regarding gender bias in academia shows that female instructors are often rated lower by students than their male counterparts. Mechanisms are complex and intersectional and often associated with role congruity theory. Little research has examined parallel patterns in graduate teaching assistant (TA) evaluations. In research institutions, TAs make up a large portion of teaching teams. Identifying bias and working to remove it is critical to shifting the already-well-documented gender imbalance in higher education. To evaluate gender-associated perceptions of graduate TAs' teaching skills, we analyzed Likert-scale, mid-semester survey data using ordinal logistic regression models for PhD TAs in five (pre-COVID) semesters in the College of Engineering at Cornell University, a large R1 institution in the United States. We also regressed scores for each survey question against the overall TA quality rating for male- and female-identifying TAs to compare the strength of those relationships and explore potential differences in student expectations associated with gender roles. A subset of narrative comment data were coded into themes, analyzed, and triangulated with other observed patterns.

Results Male TAs had a higher likelihood of receiving a better rating than female TAs for all survey questions in which students rated performance. Statistical evidence of different slopes of relationships between particular questions and overall TA quality rating suggested that female and male TAs were "valued" more for behaviors/attributes congruent with roles ascribed to that gender in broader society. Female TAs received a higher proportion of positive comments for communication skills and more comments regarding supportiveness than male TAs. Males received more comments about their overall value as TAs, however all comments regarding overall quality as TAs were positive regardless of gender. The amount and proportion of comments that were positive or negative for knowledge, enthusiasm, preparedness or fairness were the same for male and female TAs.

Conclusions Gender-based disparity is occurring in TA evaluations and aligns with patterns observed in research on teaching evaluations for faculty. Correlation between overall TA ratings and scores for specific survey questions and narrative responses indicate that role congruity influences traits that students perceive as important and positive in TAs of different genders.

Keywords Teaching evaluations, Equity, Gender, Graduate TAs, Inherent bias, Student evaluation of instruction, Role congruity theory

*Correspondence:

C. A. Evans
cae223@cornell.edu

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Introduction

Gender disparity in STEM education and employment trends are widely acknowledged and well researched (e.g., Grieco and Deitz, 2023; Hill et al., 2010; Jesse, 2006; Mastekaasa & Smeby, 2008; Modi et al., 2012; Weeden et al., 2020). A large portion of the gender disparity in higher education and subsequent employment for women in STEM continues to be associated with math, computer science, physics, and engineering. Overall, STEM enrollment in U.S. institutions of higher education between 2008 and 2018 has been consistently slightly dominated by women (55%) (Hamrick et al., 2019). However, most students graduating with bachelor's degrees in engineering from U.S. programs are men (National Center for Education Statistics, Integrated Postsecondary Education Data System [NCES & IPEDS], 2016, Fall 2015, Fall 2016).

This difference in undergraduates who successfully complete engineering curricula is a factor impacting the proportion of women who enter graduate degree programs and subsequently enter academia as instructors. In 2021, women represented 25% of undergraduate enrollment and 25% of doctoral engineering programs nationwide. According to the same source, the number of tenured or tenure track female faculty in Engineering in the United States, is approximately 19%, with variation across different engineering disciplines (American Society for Engineering Education, 2022, p. 56). Clearly there remain large challenges for women in STEM higher education and specifically, engineering. The data presented here are from the United States, however, this disparity is global in scope (Huang et al., 2020). Many documented challenges to female-identifying students and faculty in STEM disciplines have been linked to gender biases and related issues, such as stereotype threat and a “fixed” mindset that emphasizes “innate ability” over effort and growth (Blackburn, 2017; Llorens et al., 2021). These persistent attitudes influence women's self-efficacy and feeling of belonging and can decrease the desire to stay in and to advance in particular fields (Clark et al., 2021).

Implicit bias and gender bias in STEM disciplines

Implicit biases may result from what Greenwald and Banaji (1995) refer to as implicit cognition. They articulated that “the signature of implicit cognition is that traces of past experience affect some performance, even though the influential earlier experience is not remembered in the usual sense—that is, it is unavailable to self-report or introspection” (p. 5). Implicit gender biases are related to long standing institutional and societal “norms.” Biases against women in academia may be both internalized by women and imposed upon them by others in their sphere of influence across the education

landscape (Hughes et al., 2017; Weisshaar, 2017; Wittman et al., 2019). For example, faculty bias while reviewing applications for laboratory manager employment favored students that they believed were male in a double-blind study. This bias was independent of the gender of the faculty (Moss-Racusina et al., 2012). Student biases toward female instructors are also well studied (Adams et al., 2022; Aragón et al., 2023; Boring, 2017; Buser et al., 2022; Ceci et al., 2023; Chatman et al., 2022; MacNell et al., 2015; Mengel et al., 2019; Wittman, et al., 2019). Female instructors are often rated lower by students than are their male counterparts even when evidence suggests that female instructors are just as effective (Boring, 2017). Evidence for these gender-biased patterns come from both post hoc analyses of student teaching evaluations over multiple years and across disciplines (Adams et al., 2022) and experimental research designs that directly test and show differences in evaluations as biases (MacNell et al., 2015; Mengel et al., 2019). Differences in responses on evaluations for teachers who identify as male *versus* those who identify as female, are also influenced by disciplinary affiliation (Basow & Montgomery, 2005), cultural identity (Fan et al., 2019; Llorens et al., 2021), the gender of the student who is completing the evaluation (Boring et al., 2016; Fan et al., 2019), specific gender expectations for behavior (Eagly & Karau, 2002) and very often the intersectionality of multiple variables including gender (Llorens et al., 2021). These deeply engrained unidentified, or inaccurate, attitudes are linked to what people have seen enacted within gender identities in their personal experiences. Biases linked to lifetime exposure to slow-changing social roles and expectations are tenacious (Hughes et al., 2017).

Role congruity theory and gender bias in teaching evaluations

The role congruity theory (Eagly & Karau, 2002) builds off earlier work by Eagly and Karau (2002) and colleagues who coined the term *social role theory*, suggesting that the societal differences we perceive in people of different genders (M/F) come from the way in which men and women have historically been cast into specific roles in society. Role congruity theory posits that we have been socialized to expect different behaviors from women than from men (Piatek-Jimenez et al., 2018), which leads to our unconscious expectations and valuing of behaviors based on gender. Eagly and Karau (2002) hypothesized that nurturing, encouraging, communicating, and being accessible are expected positive aspects in female leaders and teachers, whereas command and control and disciplinary expertise might be expected and valued in men. These perceptions inhibit women moving into leadership positions and make it more difficult for them to be

perceived as successful in such situations (Eagly & Karau, 2002; Ritter & Yoder, 2004).

Evidence continues to mount regarding the inability of student evaluations of teaching (SET) to provide unbiased information about teaching quality effectively or equitably (Adams et al., 2022; Spooen et al., 2013; Strobe, 2020). Adams et al. (2022) evaluated gender and cultural differences in SET across 6 years and a diverse set of disciplines including science and engineering (nearly 400,000 responses). Their research examined gender and cultural biases in teaching evaluations. In the subset of their analysis that examines the effect of gender in English speaking teaching assistants (TAs), men were 1.25 times more likely to get higher scores from female students and 1.43 times more likely to get higher scores from male students.

Much less is known about gender patterns in student evaluations of TAs. Khazan et al. (2019) found no significant difference in teaching evaluation scores when a single online TA was assigned an identity of “male” for half of the students and “female” for the other half. There was however a wider range in the evaluation scores for the TA that students were told was female as well as five times as many negative comments for the TA presumed female. The authors interpreted this as an optimistic outcome, showing a lack of gender bias in evaluating TAs. However, they also cautioned that a lack of personal contact and relatively low interaction with the purported online TA may have reduced the triggering of bias. More research focused on patterns associated with gender (among other variables) perceptions of graduate TAs is clearly needed. The positionality of graduate TAs is different than that of instructors/faculty members relative to the students they teach, thus we might expect different outcomes from the large body of work examining student gender biases toward faculty, at least in magnitude. If, however, social role congruity is a strong driver in persistently male dominated disciplines like engineering, we would predict similar patterns.

A subset of PhD graduate students will become faculty of the future, and during their graduate school experience are developing skills and self-efficacy as teachers. The fact that ten percent fewer women are in tenured/tenure track faculty positions than are enrolled in engineering graduate programs means that understanding factors that influence the experience of female graduate students in engineering is critical. Any barriers to success at the graduate student level have the potential to create obstacles for female graduate students to advance in academia (Boring, 2017), specifically in this case, into faculty positions. Experiences as TAs and associated outcomes of SET are one aspect of graduate student work that deserves more attention.

This research is an examination of gender-associated patterns in five pre-COVID semesters of teaching-assistant mid-semester evaluations in the College of Engineering at Cornell University, where graduate students take on substantial roles in teaching. We specifically asked: (1) is there a difference in the “overall TA rating” (a non-specific survey question) for male *versus* female TAs by engineering students? (2) Are students likely to rate male and female TAs differently depending on the behaviors or skills that are specified in a particular survey question? And (3) do the relationships between ratings for any specific behavior or skill (survey question) and the “overall TA rating” depend on TA gender?

Methods

This research was undertaken after receiving an exemption notification (protocol ID# 2009009842) from the Institutional Review Board at the University. The aim of the work was to determine if there was evidence for gender bias in student responses on graduate (specifically, PhD) teaching evaluations and to further explore any differences in evaluation patterns between student responses regarding male and female TAs that would suggest potential mechanisms such as role congruity theory.

Gender identity data associated with TA evaluations was added prior to the dataset being anonymized. The database used for mid-semester evaluations does not currently include gender identity. The gender ID data attached to each TA in the dataset came from the University database. For reasons unknown to the authors, the choices offered to students to identify gender in this database are M (male), F (female) and U (unidentified). Surprisingly, only one TA identified as U, and three did not answer the question. Those individuals were removed from the database for these analyses. The authors note that we strongly support the reporting of a more realistic spectrum of gender identities across our student population, such as using the Gender/Sex 3×3 framework in student surveys (Beischel et al., 2023); however, based on the source of these gender data, TAs were not able to identify with a broader array of choices and thus we report only on patterns associated with TA scores for those in our population who identified as male and female. For the remainder of the paper as we share or discuss our results, we will refer to male or female TAs, or TAs who identify as male or female in this study.

Study population

The undergraduate student body in the College of Engineering is approximately 3,000 students of whom nearly 50% are women, which is a high proportion of women and continues to be rare for engineering programs.

In data analyzed for this study, male undergraduate respondents numbered 2800, and female undergraduate respondents 2742: essentially equal proportions. The PhD population continues to be male dominated and the ratio of male to female PhD TAs in the data set was 2.20 and was fairly consistent across the five semesters (Table 1). Graduate TAs work in courses offered by 18 engineering programs (Table 2) and the number of TAs working in each program varies greatly. TA roles range from lecturer, grader, recitation or discussion section leader, and

office hours instructor. Some TAs have singular roles and some multiple roles. The role of the TA was not explicitly considered in the analysis of student evaluation survey responses.

Students in other graduate programs (MS, MEng, and others) also hold teaching assistantships in the departments, as do undergraduates. We chose to focus on PhD TAs in this study because they are closest to faculty in their professional pathways and most likely to have aspiration to faculty positions. Persistent large discrepancies

Table 1 Number of male and female TAs included in the five semesters of data. Total teaching assistant (TA) numbers include those TAs that taught in more than one semester

Semester	Total Responses	Female TAs	Male TAs	Male/Female TA ratio	Female TA # responses	Male TAs # responses
FA17	1345	51	121	2.4	430	915
FA18	1092	50	116	2.3	355	737
FA19	1343	63	120	1.9	337	1006
SP18	1019	47	110	2.3	310	709
SP19	744	44	95	2.2	274	470
Totals	5543	255	562		1706	3837

The number of unique PhDTAs in the data set is 620 (432=m, 188=f).

Table 2 Number of student responses for male and female teaching assistants (TAs) in 18 programs or departments over the five semesters of mid-semester evaluation data used for this analysis

Program/ department	FA17		FA18		FA19		SP18		SP19		Study	
	TA Gender		TA Gender		TA Gender		TA Gender		TA Gender		TOTALS	
	F	M	F	M	F	M	F	M	F	M	F	M
AEP	30	17	0	55	4	47	8	29	0	14	42	162
BEE	4	0	43	12	31	13	14	2	9	8	101	35
BME	46	25	9	53	12	18	15	19	78	5	160	120
CEE	79	95	34	43	30	111	13	10	3	30	159	289
CHEME	11	117	24	42	16	76	14	24	24	18	89	277
COMM	4	4	0	0	12	7	7	0	5	2	28	13
CS	28	209	53	106	17	132	7	159	2	85	107	691
EAS	0	10	4	80	49	18	59	30	10	13	122	151
ECE	19	25	0	16	4	28	24	65	4	34	51	168
ENGRD	33	101	98	83	1	233	34	96	17	43	183	556
ENGRI	0	17	0	10	10	0	7	19	19	0	36	46
INFO	35	13	6	11	18	9	39	61	42	31	140	125
MAE	11	67	11	100	56	154	26	100	13	55	117	476
MSE	31	79	10	69	50	31	12	28	14	24	117	231
ORIE	84	107	62	50	27	112	30	50	25	87	228	406
PHYS	0	11	1	3	0	3	1	7	0	3	2	27
STSCI	0	0	0	0	0	0	0	0	9	12	9	12
SYSEN	15	18	0	4	0	14	0	10	0	6	15	52
TOTALS	430	915	355	737	337	1006	310	709	274	470	1706	3837
									Responses		5543	

in female faculty in engineering programs indicates the need to understand better the experience of women in PhD programs.

Data collection

Likert-scale, mid-semester evaluation data

Early in each semester, TA names are uploaded to a common database by administrators in each department. In the database, TA names are associated with the course in which they work. Prior to launching the mid-semester evaluations, final lists of TAs for each course are sent to the course faculty to be checked and corrected. At the mid-point of each semester (typically weeks 6–8) the survey is launched. Students receive an email with

instructions, rationale, encouragement, and a link to the survey. They are asked to select their TA(s) from a list associated with the course and to respond to 18 Likert-style questions about specific TA behaviors and practices (refer to Table 3 in results). Two additional questions asked students to (1) rate the course overall—not including the TA—and (2) rate the overall quality of their TA's teaching (referred to going forward as “overall TA rating”). Responding to the survey is voluntary and students receive at least two additional reminders to complete it during the two-week period in which it is available. Separate reminder emails are also sent to the TAs to encourage them to remind students. Students are encouraged to answer only the survey questions that related to their

Table 3 Outcome of ordered logistic regression for each survey question

Reference gender = M	TAs in analysis (n)	Estimate	Odds ratio	Lower–Upper CL	z statistic	Holms-Bonferroni adjusted p-value
1. Demonstrates command of the subject matter	610	0.424	1.529	1.314–1.743	3.878	0.0021
7. Provides clear, relevant and understandable responses to my questions	608	0.412	1.509	1.297–1.722	3.795	0.0028
16. Fair in grading	609	0.340	1.405	1.189–1.621	3.084	0.0349
20. Quality of your TA's teaching? (1 being poor, 5 being great)	590	0.335	1.398	1.203–1.593	3.372	0.0134
3. Provides clear and comprehensive explanations and instructions	608	0.333	1.394	1.172–1.617	2.930	0.0509
9. Actively helpful when students need assistance	611	0.3113	1.365	1.157–1.573	2.931	0.0541
15. Makes effective use of illustrations and examples	607	0.289	1.335	1.103–1.566	2.438	0.0888
12. Periodically checks to make sure students understand what was covered	607	0.281	1.325	1.105–1.545	2.506	0.1220
17. Provides helpful comments on my assignment	592	0.277	1.319	1.120–1.517	2.728	0.0892
13. Provides periodic summaries of what has been covered or discussed	582	0.263	1.301	1.091–1.511	2.454	0.0987
11. Communicates clearly	597	0.262	1.299	1.109–1.489	2.702	0.0896
4. Emphasizes the conceptual basis of the problem set or the lab experiment	597	0.258	1.294	1.092–1.496	2.502	0.1116
18. Makes effective use of visual aides (blackboards, overhead, slides etc.)	583	0.253	1.288	1.091–1.485	2.518	0.1298
14. Effective at relating lecture material to what is covered in section or lab	589	0.249	1.283	1.095–1.471	2.600	0.1118
10. Seems enthusiastic about teaching the material	595	0.247	1.280	1.080–1.479	2.416	0.0785
5. Encourages students to think in class by asking questions	608	0.232	1.261	1.043–1.478	2.088	0.1472
6. Makes me feel free to ask questions and express my opinions	581	0.231	1.260	1.077–1.442	2.481	0.1048
19. Divides his/her time equitably among laboratory groups	565	0.2066	1.229	1.033–1.426	2.057	0.1191
2. Fully prepared for class, laboratory or review section	603	0.182	1.199	0.975–1.424	1.587	0.2260
8. Evaluate this course as a whole. (1 being poor, 5 being great)	614	0.074	1.077	0.890–1.264	0.780	0.4350

The Odds ratio is interpreted as: the odds that a male TA has a higher response level is (the Odds Ratio) times higher than it is for a female TA in our data set. 95% confidence intervals are also reported. Bonferroni-Holms post-hoc adjustment was applied to correct for multiple comparisons in the interpretation of statistical significance. Data were sorted so that odds ratios are listed from largest to smallest

*Holms-Bonferroni corrected α used to correct for multiple comparisons. Data were sorted by unadjusted p-value (smallest to largest) and the corrected p-value was determined by multiplying the unadjusted p-value by total number of comparisons (20) for the most significant outcome and multiplying each successive p-value by one less comparison for each subsequent correction. Adjusted p-values ≤ 0.05 are in **bold**

interaction with the TA. For example, if a TA is not a grader for students, rather held office hours and recitations, a student should leave the question regarding “fair grading” blank. Alternatively, if the only relationship a student has with a TA is through the grading of exams and problem sets, they should only respond to questions about “fair grading” and “feedback on assignments.” Due to this, the sample size of responses for each survey question were slightly different.

Because of concerns over survey burnout, the TA evaluation survey is administered only once as a formative instrument. TAs (and their faculty) receive their scores with suggestions on how to interpret them and how best to use them to make changes, if necessary, during the rest of the semester. Students are asked to evaluate their professors’ teaching in separate surveys at mid-semester and at the end of the semester.

Data for this analysis come from five pre-COVID semesters of TA mid-semester evaluation data (Fall 2017, Spring 2018, Fall 2018, Spring 2019, Fall 2019). Pre-COVID semesters were chosen due to the emergency online teaching that was implemented during the Spring 2020 to Spring 2022 semester for at least a part of each of those semesters. The effect of emergency online teaching with inexperienced faculty and TAs was not a variable in which we were interested. We examined anonymized data for patterns associated with male and female TAs’ scores for Likert-scale survey questions and a subset of narrative comment data, collected from students in all engineering courses in which they interact with TAs, including through grading.

Narrative data

As part of the evaluations, students respond to three narrative evaluation prompts. Two are completely open ended (“Comment on the TA’s teaching strengths as well as areas in which improvement is needed or encouraged” and “Do you have any additional comments?”), and one is more focused and invites students to focus on communication in general, and language specifically (“Comment on the TA’s communication strategies. Did the TA effectively use gestures, movement, voice inflection, and maintain eye contact? Was language a barrier to your understanding? If so, in what way?”). A random number generator was used to select a subset of 472 students’ comments (in proportion with M/F TA ratio in the dataset) for TAs whose average overall TA rating score was a 3, 4, or 5. The choice to exclude TAs with an overall TA rating score lower than 3 (on a Likert-scale of 1–5) was made to remove the influence of potentially large differences in TA quality from the examination of comments for TAs based on their gender. We supposed that TAs with extremely low ordinal response ratings would

be more likely to have comments that said less about skills and behaviors and more about student frustration regardless of TA gender.

Statistical analysis

Ordinal logistic regression

We used regression models for ordinal data. (Christensen, 2022, Ordinal package version 2022.11-16) in R (R Core Team, 2022) statistical software to analyze the effect of gender on responses for each of the 18 Likert-scale questions regarding specific skills or practices, one question rating a general notion of “overall TA rating,” and one additional question related to the course overall ($n=20$ survey questions total; 5543 total student responses for 620 PhD TAs, $n=432$ male TAs, and $n=188$ female TAs). Some TAs taught in multiple semesters; thus, individual TA ID was included as a random effect in the models. We applied a Holms-Bonferroni correction for multiple comparisons to address concerns about the likelihood of inflated type 1 error (Wright, 1992). We note here that the discussion surrounding how to deal with multiple comparisons is important and expert opinions are highly variable. Correcting for a large number of multiple comparisons can inflate the chance of a type 2 error (Midway et al., 2020). P -values with $\alpha \leq 0.05$ are typically chosen to represent statistical significance and we use that convention here. However, correcting for a large number of comparisons, such as we examined, greatly increases the p -values of those individual tests. As noted by Wasserstein et al. (2019), p -values are one way to ascribe “importance” to outcomes, and statistical significance is not necessarily equal to contextual importance. We focus our discussion and conclusions on all the forms of evidence provided in our analyses.

Relationships between individual questions and overall TA rating

Spearman correlation To explore the degree to which individual skills/behaviors were correlated to perceived performance in TAs of different genders, Spearman correlation (JMP version 16) was used to evaluate the strength of relationship between each of the nineteen questions (this included the course rating) and the overall TA rating for TAs of each gender. Due to the ordinal nature of the data, we considered Spearman’s Rho the most appropriate analysis. We used each individual Likert response for specific questions in the analysis. Sample size for each correlation varied due to students answering only the questions relevant to their experience with TAs (refer to Table 3 in results).

Examining interactions between question scores and gender on overall TA rating

Because we were also interested in the potential that student responses to questions about particular skills/behaviors may interact with TA gender as predicted by role congruity theory, we averaged question scores for each TA and used least squares linear regression models to look for significant interaction terms between TA gender and the TAs mean score for each specific question and the TAs mean score for overall TA rating. We justify this parametric regression approach for exploring the interactions because the relationships, in all cases, were monotonic and appropriately linear (based on residual *versus* normal quantile plots), and plots of residuals *versus* predicted values were evenly distributed around the best fit line.

Narrative data analysis

We analyzed narrative comment data, for a subset of unique PhD TAs ($n=324$ comments for male TAs, $n=148$ comments for female TAs) whose overall TA rating was a 3, 4, or 5. One narrative comment was selected randomly to represent feedback for each TA in this subset. Comments for male and female TAs were 69% and 31% respectively, in proportion to the general ratio of male to female TAs. We used a “theoretically driven inductive approach” to coding the data as described in Syed and Nelson (2015, p. 4). The random sample of narrative entries that accompanied Likert-scale responses were split and coded by two independent researchers using an open coding approach in which themes emerged from the comments written by students (Creswell & Creswell, 2022). The unit of analysis was phrases long enough to clarify the context of the text associated with each theme. After independently identifying categories of emergent themes, the two researchers discussed and came to agreement through discussion moderated by a third researcher. Subsequent discussions to refine themes was informed by the theoretical underpinnings of the research focus: gender bias and role congruity theory. Finally, to acknowledge and avoid potential concerns about the potential for the theoretical framework of the research to influence coding and subsequent thematic analysis, all language that might indicate TA gender was removed from the comments and the third researcher coded the full data set independently using the agreed upon themes. Upon completion of coding the de-gendered data set, outcomes were compared with the two initial coders determinations, small discrepancies discussed, and full agreement achieved.

Students’ comments were observed to fall into eleven categories which were ultimately similar to those determined in other teaching evaluation studies (e.g., Sprague

& Massoni, 2005): communication of content, supportiveness, verbal/written skills, general TA quality, knowledge, pedagogy, enthusiasm, preparedness, fairness, confidence, and humor. Each TA’s comments were considered holistically and marked as positive or negative within each relevant category. Comments about a TA’s ability to explain things clearly were coded as “communication of content,” while comments about a TA’s helpfulness were recorded as “supportiveness.” References to methods like using guiding questions would fall under the “pedagogy” category. As an example, the following comment was given for a TA: “[TA] is incredibly enthusiastic and knowledgeable. [TA] is good at communicating the material and answering questions. Lab is generally a pleasant place to be.” This comment would be marked positively in the “enthusiasm,” “knowledge,” “communication of content,” and “supportiveness” categories. A comment such as: “[TA] demonstrates an extensive grasp of the concept material. TA does not, however, explain these concepts well. In their attempts to not give away answers, their responses end up causing more confusion than help” would be scored positively for “knowledge” and negatively for “communication of content.”

Contingency analysis was used to determine if differences between the number of comments in each emergent theme, and the direction (positive or negative) of the comments associated with each theme were differentially associated with TA gender. We chose to report Fishers Exact p-values (for small sample sizes) in all cases, even where the number of comments were enough to meet assumptions of the Chi-Square. Holms-Bonferroni correction for multiple comparisons was applied to adjust p-values for multiple comparisons.

Results

Gender patterns in survey question scores—logistic regression outcomes

Male TAs had a 40% increase in the odds of receiving a higher score than female TAs for the general question about overall TA rating (Odds ratio (OR)=1.405, 95% CI [1.189–1.621]) (Table 3). Male TAs were also approximately one and a half times more likely to receive higher scores for students’ perception of the “command of subject matter” (OR=1.53, 95% CI [1.314–1.743]), “providing clear, relevant and understandable answers to questions” (OR=1.51, 95% CI [1.297–1.722]). There was no evidence to suggest a difference between male and female TA ratings with respect to being “prepared for class, laboratory, discussion section” (OR=1.199, 95% CI [0.975–1.424]) and the question regarding the “quality of the course overall” (OR=1.077, 95% CI [0.890–1.264]). For the remainder of the questions, male TAs were

between 1.39 to 1.23 times more likely to receive higher evaluation scores than female TAs. (Table 3).

Spearman correlations between specific mid-semester evaluation questions and the overall TA rating

The correlation between TA scores on specific tasks/responsibilities and overall TA rating may be an indication of different expectations and values that students place on skills that TAs are expected to perform. Spearman correlations coefficients between questions and overall TA rating tend to be higher for female TAs than male TAs in all but one question (“divides time equitably”) (Table 4). This outcome suggests that specific perceived skills (as measured by survey questions) for female TAs are more indicative of the student’s perception of the value of female TAs overall.

The strength of the correlations can generally be interpreted as an indication of the importance to students of

particular skills in TAs. While there is a trend of higher correlations for female TAs in general, skills that appear to be most important to students regardless of TA gender are: clear, relevant responses; comprehensive explanations; clear communication; and effectively relating the work in the TA-led session to the material covered in lecture (Table 4).

Interactions between individual question scores and gender on overall TA rating

We included this analysis in our mixed methods approach to explore the data for evidence consistent with role congruity theory. Table 5 includes regression analyses’ output for the five survey questions for which there was statistical evidence for an interaction between the two independent variables (the “question” and “TA gender”) on the response variable (overall TA rating). Interactions are represented by differences in the slopes of

Table 4 Spearman correlations between individual survey questions and “overall teaching assistant (TA) rating” for male and female PhD TAs. Questions are sorted from highest to lowest correlation coefficient for each gender

Survey Question: “My TA,” or “My TA is:”	Female TA r (n)	Survey Question: “My TA,” or “My TA is:”	Male TA r (n)
Provides clear, relevant and understandable responses to my questions	0.83 (1516)	Provides clear, relevant and understandable responses to my questions	0.78 (3463)
Provides clear and comprehensive explanations and instructions	0.82 (1542)	Provides clear and comprehensive explanations and instructions	0.77 (3508)
Effective at relating lecture material to what is covered in section or lab	0.79 (1341)	Communicates clearly	0.76 (3535)
Communicates clearly	0.78 (1554)	Effective at relating lecture material to what is covered in section or lab	0.75 (3138)
Emphasizes the conceptual basis of the problem set or the lab experiment	0.76 (1440)	Emphasizes the conceptual basis of the problem set or the lab experiment	0.72 (3338)
Periodically checks to make sure students understand what was covered	0.74 (1466)	Actively helpful when students need assistance	0.71 (3517)
Makes effective use of illustrations and examples	0.74 (1427)	Makes effective use of illustrations and examples	0.71 (3251)
Actively helpful when students need assistance	0.72 (1550)	Seems enthusiastic about teaching the material	0.69 (3496)
Seems enthusiastic about teaching the material	0.72 (1547)	Divides his/her time equitably among laboratory groups	0.69 (2629)
Provides periodic summaries of what has been covered or discussed	0.72 (1406)	Demonstrates command of subject matter	0.68 (3534)
Makes effective use of visual aides (blackboards, overhead, slides etc.)	0.72 (1343)	Periodically checks to make sure students understand what was covered	0.68 (3326)
Demonstrates command of subject matter	0.71 (1557)	Provides periodic summaries of what has been covered or discussed	0.68 (3181)
Encourages students to think in class by asking questions	0.71 (1411)	Makes effective use of visual aides (blackboards, overhead, slides etc.)	0.68 (3039)
Fully prepared for class, laboratory or review section	0.70 (1491)	Provides helpful comments on my assignment	0.68 (2821)
Makes me feel free to ask questions and express my opinions	0.69 (1523)	Makes me feel free to ask questions and express my opinions	0.66 (3467)
Provides helpful comments on my assignment	0.69 (1281)	Fully prepared for class, laboratory or review section	0.66 (3393)
Divides his/her time equitably among laboratory groups	0.69 (1147)	Encourages students to think in class by asking questions	0.65 (3212)
Fair in grading	0.65 (1297)	Fair in grading	0.64 (2852)
Evaluate this course as a whole	0.52 (1584)	Evaluate this course as a whole	0.48 (3589)

The characteristics included in these survey questions tend toward supportive, nurturing, and structural organization types of behaviors

Table 5 Least squares regression model outcomes to examine the interaction between specific questions and teaching assistant (TA) gender as it relates to “overall TA rating” (“intercept” and “survey question” variables not shown). Only survey questions with evidence of clear interaction terms were included

Survey Question	Variable	Coefficient	St. Err	p-value
Provides clear and comprehensive explanations and instructions	Gender	0.00097	0.0145	0.9463
	Question*Gender [F]	− 0.0767	0.0226	0.0007
Provides clear, relevant and understandable responses to my questions	Gender	0.86104	0.02058	<0.0001
	Question*Gender [F]	− 0.04549	0.02058	0.0274
Periodically checks to make sure students understand what was covered	Gender	−0.02705	0.01538	0.0793
	Question*Gender [F]	0.09373	0.02404	0.0001
Makes effective use of visual aides (blackboards, overhead, slides etc.)	Gender	−0.04169	0.01823	0.0226
	Question*Gender [F]	0.10136	0.03182	0.0015
Provides periodic summaries of what has been covered or discussed	Gender	−0.0291	0.01666	0.0811
	Question*Gender [F]	0.06295	0.02548	0.0138

relationships between male and female TAs for a specific question *versus* overall TA rating. These results suggest that male TAs were more valued overall when they provided “clear and comprehensive explanations,” and “clear, relevant and understandable responses to questions” (Fig. 1a, b). Note that in Fig. 1, the slopes of the best fit lines are steeper, and the intercepts are lower for male TAs than for female TAs. We interpret this as suggesting that male TAs perceived as having low levels of “knowledge/expertise” are more likely to receive an overall lower score than female TAs with the similar low scores on the same question. Using the same interpretation, female TAs were valued more “overall” compared to male TAs if they were perceived to “check to make sure that students understand what was covered,” “provided periodic summaries of what was covered,” and “make effective use of visual aids (blackboards, overheads, slides, etc.)” (Fig. 2a–c).

Narrative analysis of comments for male and female TAs who received 3,4, or 5 for overall TA rating

Of the three narrative response questions asked in the mid-semester evaluations, two were open-ended and one was related to communication and language use. Thus, it is not surprising that the largest number of comments were related to communication and many of those were related to verbal and written skill (40% of comments coded). Responses to the open-ended questions were responsible for approximately 60% of the comments (Table 6).

Of all comments about “communication of content,” female TAs received a higher proportion than male TAs, and a higher proportion of communication comments were positive for female than for male TAs. There is evidence that female TAs also received more comments associated with supportiveness; however, there were no

obvious differences in the proportion of positive and negative comments for female and male TAs in this category. Male TAs received more comments regarding general value/quality as a TA. However, comments for all TAs in this category were positive regardless of gender. In all other comment categories, male and female TAs received approximately equal proportions and there were no differences in positive *versus* negative remarks (Table 6).

Discussion

These outcomes shed more light on the complexity of human perceptions as influenced by traditional social roles and constructs that perpetuate inequity in institutions including higher education. Assumptions, expectations, and beliefs result in implicit/explicit biases and perpetrate barriers to advancement of under-represented populations in certain fields. We focused here on the role of TA gender in student perceptions of their skills and their overall quality as educators in a college of engineering at a large R1 university. The context of the work is the persistent gap of female students in graduate student roles and subsequent faculty appointments in STEM disciplines, recognizing engineering programs as one of the largest gaps. Our results show that gender biases, similar to those shown in female faculty members’ teaching evaluations (Adams et al., 2022; Boring, 2017; Boring et al., 2016; Fan et al., 2019), exist in this earlier iteration of female instructors in higher education: PhD TAs. Student perceptions of the performance of TAs are influenced by the roles they see enacted in the larger societal context—women in supportive and nurturing roles, men in roles of knowledgeable and credible sources for answers (Adams et al., 2022; Eagly & Karau, 2002). In each of the following sections we evaluate the questions that framed the analysis of five semesters of TA mid-semester evaluation data. By triangulating the outcomes, we articulate the interplay

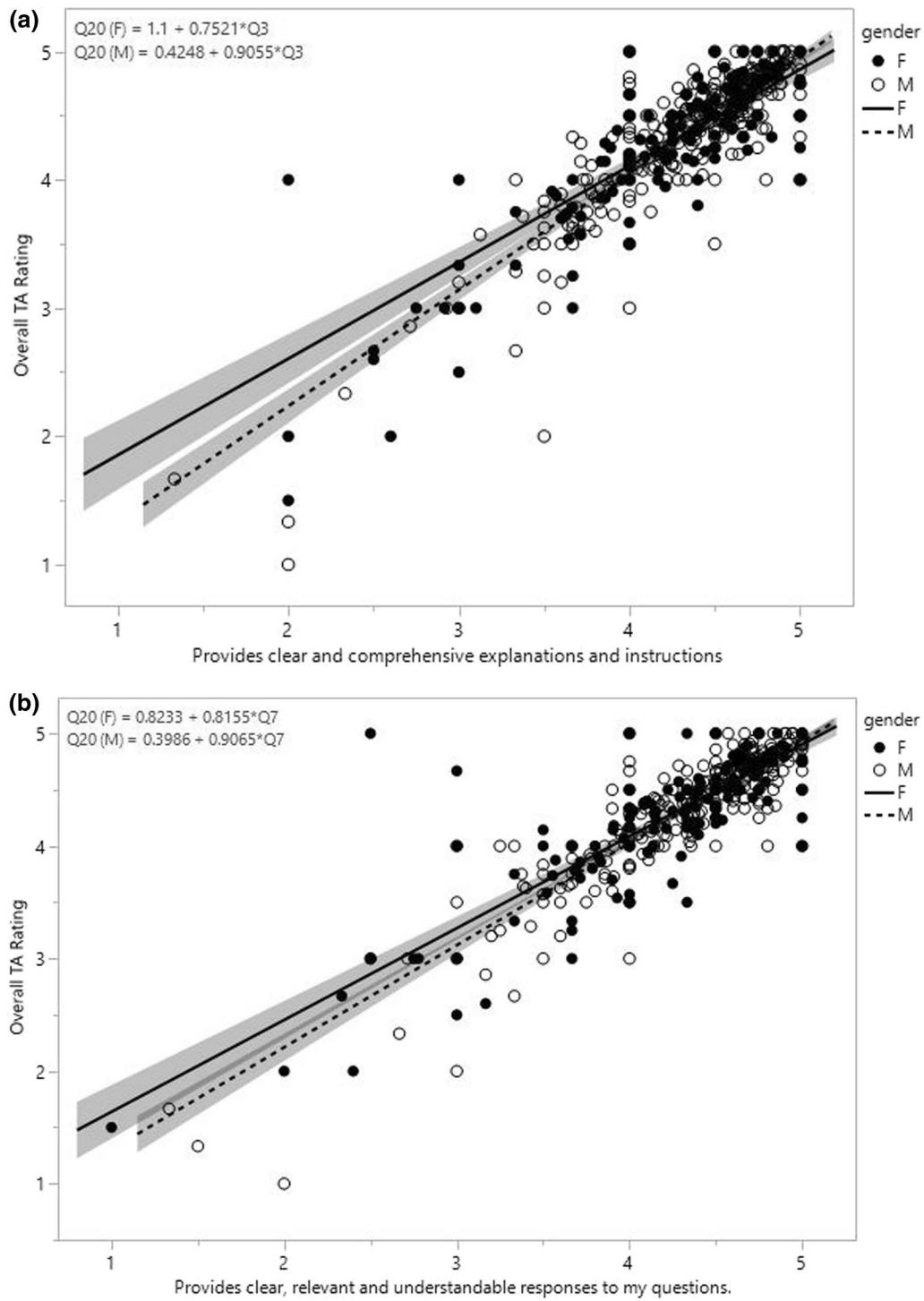


Fig. 1 a, b Relationships between survey questions and overall TA Rating (by gender) for which male overall TA ratings are lower than female overall TA ratings when male TAs are perceived as lacking in these specific skills—**a** provides clear and comprehensive explanations and instructions, **b** provides clear, relevant and understandable responses to my questions

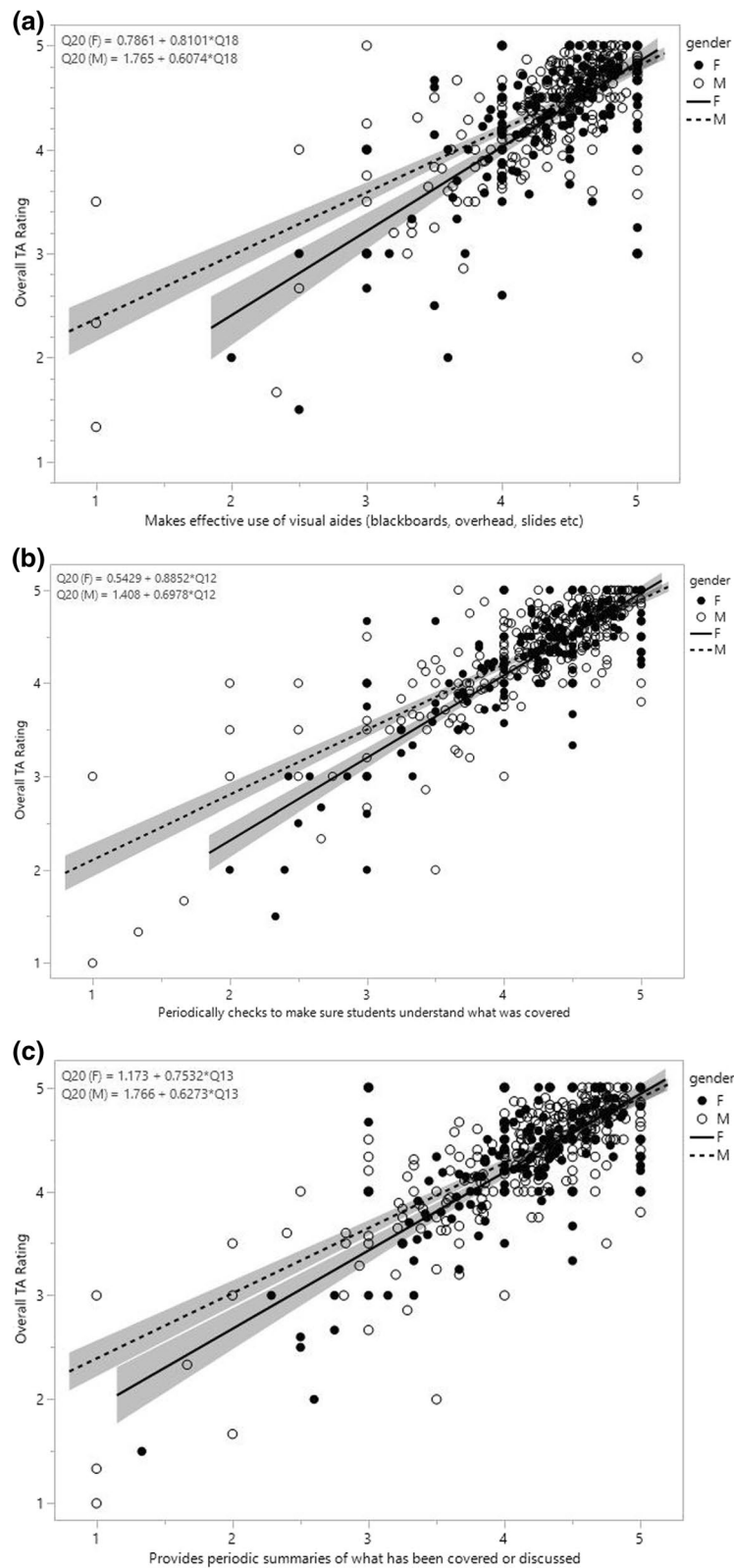


Fig. 2 a–c Relationships between survey questions and overall TA Rating (by gender) for which female overall TA ratings are lower than male overall TA ratings when female TAs are perceived as lacking in these specific skills—**a** makes effective use of visual aids **b** periodically checks to make sure students understand what was covered, **c** provides periodic summaries of what has been covered or discussed

Table 6 Narrative analysis of 472 randomly selected comments on the performance of the same number of teaching assistants (TAs) ($n = 324$ comments for male TAs, $n = 148$ comments for female TAs)

Comment Theme (total #of comments in data subset)	% Male TAs received comments	% Female TAs received comments	Fishers Exact p (2-tailed) Male vs Female comments*	% Positive Comments for Male TAs	% Negative Comments for Male TAs	%Positive Comments for Female TAs	%Negative Comments for Female TAs	Fishers Exact p (2-tailed) Male vs. Female (pos vs.neg)*
Communication of content (306)	62.04	70.27	**0.0203	80.09	18.90	89.42	8.65	****0.036
Supportiveness (228)	44.14	56.76	***0.023	92.30	7.69	90.36	9.52	0.366
Verbal/written skills (152)	32.45	27.70	0.364	67.57	32.43	79.73	29.27	0.845
General TA quality (126)	29.32	20.95	0.060	100.00	0.00	100.00	0.00	1.000
Knowledge (116)	23.77	25.68	0.786	96.10	3.90	94.73	5.26	1.000
Pedagogy (85)	19.44	14.86	0.192	60.32	39.68	77.27	22.72	0.199
Enthusiasm (56)	10.18	15.54	0.177	78.79	21.21	86.96	13.04	0.500
Preparedness (50)	10.80	10.13	0.317	57.17	42.86	33.33	66.66	0.217
Fairness (18)	4.32	2.70	0.785	64.29	35.71	50.00	50.00	1.000
Confidence (8)	1.23	2.70	0.360	0.00	100.00	25.00	75.00	1.000
Fun/humor (4)	0.93	0.68	1.000	100.00	0.00	100.00	0.00	1.000

Note that comment themes are listed in order of number of comments in each category. Percent negative and positive comments for male and female TAs are reported in proportion to the total number of male and female TAs in the narrative sample. Uncorrected Fishers Exact values are reported. Holms-Bonferroni correction was applied. See footnote

*When Holms-Bonferroni correction for multiple comparisons is applied to uncorrected Fishers exact p-values (< 0.05), none of the corrected comparisons meet the $p < 0.05$ criteria for statistical significance. Corrected values for comparison: **0.223, ***0.230, ****0.396

between overall gender bias and student expectations of behaviors in female *versus* male TAs that appears to be related to the framework of social role congruity (Eagly & Karau, 2002).

We did not use respondent gender as a variable in our analyses. We note that research regarding the effect of respondent gender on teaching evaluations is mixed. Some studies have shown trends in which students tend to rate instructors or TAs of their own gender lower than the alternate gender (Boring et al., 2016; Khazan et al., 2019), while others have shown that each gender may rate similarly-gendered instructors higher than the alternate (Young et al., 2009). Some studies show no trend or main effect of student gender on the outcomes of instructor evaluations (Fan et al., 2019). In our data, the total number of responses were evenly split between male and female students (only 58 more male than female respondents in 5542 total responses). We did not determine if male or female TAs received very different amounts of responses from either male or female students and thus cannot be sure that there was no effect of respondent gender. Regardless, the implications of bias in teaching evaluations remain the

same in the context of hiring decisions and effects on self-efficacy.

Greater likelihood of higher ratings for male *versus* female PhD TAs

Our results corroborate other research that has shown that students are more likely to rate female instructors lower than male instructors (Adams et al., 2022; Boring, 2017; Ceci et al., 2023; MacNell et al., 2015; Mengel et al., 2019; Mitchell & Martin, 2020; Witteman, et al., 2019). In a previous study examining this phenomenon in TAs, Khazan et al. (2019) did not see a significant difference in evaluation scores in an online course with a single TA who some students believed was male and others believed was female. There was, however, a broader range in the evaluation scores and many more negative comments for the purported female TA. The authors suggested that distance between teacher/TA and student in an online class may have reduced gender bias.

In our study, male PhD TAs had much higher odds of being rated more highly than female PhD TAs on all mid-semester evaluation questions regarding skills/behaviors. Also congruent with other work, men had

the highest odds ratios with respect to being perceived as knowledgeable and providing relevant information to students in clear, concise ways (Adams et al., 2022; Borning, 2017; Buser et al., 2022; Chatman et al., 2022; Young et al., 2009). The two questions that showed no apparent gender difference were “preparedness for teaching” and “evaluating the course as a whole.” It is curious that female TAs were perceived, overall, as equally prepared for their teaching roles, yet male TAs were substantially more likely to be rated highly at “demonstrating command of the subject matter,” “providing clear, relevant and understandable responses to questions,” and “providing clear comprehensive explanations and instructions”—all of which, in some part, should follow from preparedness. The fact that students’ perception of the value of the “course as a whole” was not influenced by the gender of the TA provides at least a check on students reading the survey questions and distinguishing between a question regarding the course *versus* those about the work done by TAs.

Building the case for gender bias—the role of narrative data

In a post-hoc study such as this, additional evidence is required to substantiate the claim that consistently higher odds that male TAs were perceived as doing a better job represents gender bias, rather than an exposure of the lower quality of female PhD TAs. Thus, we contrast quantitative outcomes with the narrative analysis to further build the argument for gender bias.

Female TAs not only received more comments related to communication of content than did male TAs, but a higher percent of these were positive. This is in juxtaposition to the logistic regression analysis outcomes for three survey questions associated with content communication: “communication skills,” “provides clear, relevant and understandable responses to my questions,” and “provides clear and comprehensive explanations and instructions.” Odds ratios for male TAs range between 1.33 and 1.50 for all these survey questions. Female TAs also received more comments in the category of “supportiveness” of which 90% were positive. The idea of supportiveness is reflected in a large subset of the Likert-scale survey questions, all of which had more favorable odds for male TAs. The only substantively greater number of narrative comments for men was in the “overall TA quality category”—a non-specific ranking. This parallels the higher odds that a male TA would receive a higher overall TA score. However, there are no other differences between male and female TAs in either the number of comments or the proportion of those comments that are positive or negative for more specific categories. If male TAs were actually “better” in these areas, we would also

expect narrative comment data to corroborate the outcomes of the logistic regressions in the categories related to other survey questions like “knowledge,” “verbal/written skills,” and/or “enthusiasm” for which the odds ratios for male TAs were also increase compared to female TAs, which they did not.

From the combined outcomes of the logistic regression and the narrative data analysis we find the initial evidence for bias against female-identifying PhD TAs. The excess narrative comments regarding support and communication for female TAs also indicate that students may have gendered expectations for TAs. The remainder of the discussion explores evidence suggesting that, in addition to a general bias against female engineering PhD TAs in evaluations, survey results for both female and male TAs are influenced by persistent perceptions of appropriate/expected social behaviors.

Gendered expectations for TAs—Spearman correlations for specific skills *versus* overall TA rating

We interpret the relationship between individual question scores and overall TA rating as (1) an indication of which behavior/skills students considered most important, and (2) those behavior/skills expected based on TA gender. In general, students strongly valued clear and relevant responses to their questions and comprehensive explanations and instructions, contextualizing learning, and relating the topics addressed in TAs’ sessions back to lecture materials. Least valued skills overall were providing helpful comments on assignments, making students free to ask questions and express opinions, and fairness in grading.

Contrary to our results, the influence of perceived grading fairness has been shown to have at least a moderate effect on student teaching evaluations for faculty (Griffin, 2004; Marks, 2000; Spooen et al., 2013). Griffin (2004) reported that students who did less well than they expected or deemed instructors as non-lenient in grading practices rated instructors lower on evaluation questions. Our TA evaluations are intentionally formative and occur mid-semester, so it is likely that students are less focused on the grade they will ultimately receive in the course. Additionally, graduate students typically do not have control over final scores that students receive, which could be another reason that repercussions for TA ratings associated with grading are reduced in the present study.

Spearman correlation coefficients for TA skills ratings *versus* overall TA rating varied in both strength and rank-order based on gender, and in most cases, the correlation coefficients trended higher for female TAs as compared to male TAs (not statistically analyzed). Most notable differences in rank-order of importance (correlation strength) of particular skills included higher ranking for

female TAs for “periodically checking on student understanding” and “encouraging students to think in class by asking questions” and for male TAs for “dividing time equitably among laboratory groups.”

Stronger correlations between specific survey questions and the overall rating indicate that the overall TA rating is perhaps less arbitrary and more a function of a set of perceived specific skills. The trend of higher correlation between skills and overall TA ratings for female TAs may suggest that their (within gender) mid-semester evaluation scores more honestly reflect the implementation of practices for which we are seeking mid-semester feedback than do scores for male TAs. Less scrutiny of male TA skills due to a general bias in their favor could lead to more variability in the correlations as well as make the feedback less valuable as a formative tool—its intended use.

Evidence for role congruity theory—relationships between ratings for specific behaviors/skills and overall TA rating by TA gender

Several studies have confirmed that student expectations for female faculty emphasize skills and behaviors that are both more time consuming and emotionally intensive than their male counterparts (Boring, 2017; MacNell et al., 2015; Sprague & Massoni, 2005). Recently, Adams et al. (2022) confirmed that social role congruity influenced faculty teaching evaluations and concluded that female faculty were expected by students to do more of the socially demanding tasks. We posited that if social role congruity were influencing student responses on TA teaching evaluations, the slope of the regression lines for average score *vs.* average overall rating score for some survey question would be significantly different for male and female TAs. Results showed that female TAs (as a whole) received disproportionately low overall ratings if they were perceived to be deficient in skills related to supportive tasks like checking understanding, occasionally providing summaries of topics or materials, and organizational skills like using effective visuals. Note that, like previous outcomes for female faculty, these tasks/skills require more planning and preparation. Male TAs who were perceived to be lacking the ability to give clear relevant responses to questions and comprehensive explanations and instructions were more likely to receive disproportionately low overall TA ratings. In contrast to expectations for female TAs, these skills valued in male TAs are centered around conveying knowledge and direct communication in the moment and require less preparation and planning.

In our data set, students generally valued practices and skills that were more teacher-centered (i.e., clear, relevant responses, and comprehensive explanations and

instructions) than those more student-centered practices that have been shown to be most effective for promoting knowledge generation and critical thinking (providing feedback on assignments, asking questions to encourage thought, and creating an atmosphere where student share opinions and are comfortable collaborating and discussing) (Connell et al., 2016; Freeman et al., 2014; Smith et al., 2014). If students place more value on teacher-centered pedagogy, and they show bias in attributing these specific characteristics to male TAs, female TAs will be even more disadvantaged in survey outcomes.

Limitations of this study are those that apply to any post-hoc analysis that uses mixed methods to build evidence and make arguments. The mechanisms and meanings ascribed to correlation coefficients, interactions between the nature of specific survey questions and gender of TA, and the juxtapositions between quantitative and qualitative analysis used to triangulate and draw conclusions are ultimately open to interpretation. In addition, as mentioned, the binary gender designation was used because the dataset from which these data were obtained is antiquated. Graduate students in this database were only offered the choices of M and F (or U). Likely because of this, most TAs identified as M or F. We view this as a limitation to this study. We do not subscribe to binary gender identities and hope that a full slate of gender identities will be available for more thorough analysis in future work. We also note that our outcomes are generalized across TAs who are engaging in varying types of teaching tasks, and also who have varying levels of teaching experience. For future studies, it could be particularly valuable to know how these variables influence student perceptions of TA skills by gender. Finally, with interactions between general bias against female TAs and bias ascribed to social role congruity in our outcomes, it is challenging to articulate effect sizes. Regardless of effect size, with the barriers faced by women in engineering programs, any effect is too great.

Intersections, challenges, and moving toward change

Creating change and equitable opportunities for women and other underrepresented populations in higher education is hard and complex, particularly in disciplines that have been historically, and are still, dominated by men. The work of Armstrong and Jovanovic (2015) reminds us that challenges faced by women in higher education, while they can be studied one variable at a time, are multivariate and the more detailed information that occurs at the intersection of gender, race, and culture needs more attention. Solutions to increase the number of women who are interested and successful in STEM programs can be categorized into “supply-side” and “demand-side” interventions (Hughes et al., 2017; Salmon, 2022).

Programs to address supply-side challenges are prevalent and include providing high school and undergraduate female students with programs that build confidence, competency, and interest in STEM fields where their numbers are low. These programs can be successful in developing identities and skills as math and science learners and work on the side of getting and keeping more members of underrepresented groups in the “pipeline” (Akin et al., 2022; Hunt et al., 2021). Demand-side challenges are those obstacles to women’s successful navigation of and ability to thrive in these roles in the environments where roles are enacted. Research that uses the pipeline analogy tends to refer to these sorts of obstacles as “leaks” in the pipeline (Almukhambetova et al., 2023; and many others). The work shared here is representative of the demand-side challenges that are a part of persistent social, institutional, and disciplinary biases. The authors of this work believe there must be a distinction made between what we refer to as “leaks” in the pipeline for women in science and engineering and the barriers created by implicit and explicit biases. “Leaks” suggest a voluntary or passive leaving of women from STEM pathways. The persistent biases that are constantly documented, including this work, amount to more than passive losses of women with aspirations and talent being excluded from their chosen work. Women rather are often being squeezed out of this work because of pressure from “barriers” or “roadblocks” to persistence and advancement, all along the “pipeline.”

Studies that corroborate social role congruity theory are disheartening because they expose a manifestation of long-entrenched phenomena that permeate all aspects of human experience and, as such, are not “fixable” in the short term. Even in the few institutions where undergraduate populations are now close to gender parity (as they are at the institution from which these data come), substantial disparities in the proportion of female and male graduate students (and faculty) persist. Demographics of female graduate students and female faculty in tenure track positions suggest that these demand-side challenges are not going away any time soon (Hughes et al., 2017).

Several studies have suggested that this form of anonymous online evaluation should no longer be used for hiring or promotion decisions or, at the very least, should be supplemented by other feedback mechanisms such as focus groups, portfolios, peer observations, and interviews (Adams et al., 2022; Baldwin & Blattner, 2003; Lattuca & Domagal-Goldman, 2007). Peterson et al. (2019) showed that simply reminding students about the implicit biases against female faculty and inviting reflection on these prior to beginning the survey removed gender bias. Students in the treatment groups in the study were

reminded specifically about the prevalence of gender bias on evaluations. The intervention improved the ratings for female faculty ($n=2$) but did not change the ratings for male faculty ($n=2$). We agree that reminding students about biases at the beginning of any teaching evaluation should be a regular practice if we continue to use this form of feedback to improve teaching effectiveness. As Peterson et al. (2019) pointed out, while the positive shift in scores for female faculty were important outcomes of that study, it was impossible to know if respondents were overcompensating for the known gender bias against women. Considering the multidimensionality of implicit biases, we suggest that the administration of surveys include not only considerations of gender but all potential biases that a person might have based on their own lived experience. Crafting a statement that invites consideration of one’s unconscious biases should include reminders about how our perceptions of gender roles, gender identity, race and cultural, and perhaps other (i.e., age) differences have been shown to influence feedback in ways that bias outcomes and defeat the purpose of the process (Chatman et al., 2022).

We also suggest making survey questions as specific as possible so that students are required to focus their evaluation on more specific tasks and pedagogical choices related to evidence supported teaching in the classroom. As we have shown, asking about specific, tractable behaviors associated with evidence supported practices will not eliminate biases informed by students’ personal indoctrination about gender roles in society, but a focus on specific skills can at least help to direct student reflection.

Finally, requiring students to answer semi-structured narrative questions before seeing and answering ordinal types of questions would allow for reflection and the generation of examples before viewing the Likert-scale portion of the survey. If students have retrieved memories of experiences interacting with their TAs, they may have more reflection with which to better focus on answering the specific question with more clarity.

Conclusions

This work examining PhD teaching-assistant evaluations corroborates gender patterns found in student expectations and biases related to faculty teaching evaluations and informs us that, not surprisingly, gender biases for female educators in higher education do not begin when they advance to faculty status, but in fact are likely one of the many reasons relatively few women do. Our analyses showed a general bias against female TAs and also more complex influences of student perception of gender roles that corroborate social role congruity theory in teaching evaluations. Female TAs tended to receive lower overall TA rating scores if they were perceived as unsupportive

or if they did not make teaching accessible by checking understanding or providing good power points. Male TAs tended to score lower overall if they were seen as not providing clear, relevant, and comprehensive answers and instructions. Many studies' authors have concluded that SET contain biases and should not be used as evidence in tenure and promotion or hiring. Evaluations remain one of the easiest metrics with which to provide feedback on teaching practices, and so are hard to let go. If they are used, they should be used as formative measures for improvement of a TAs personal teaching practice. Even in this case, inviting respondents to consider their complex biases, inviting students to reflect on semi-structured narrative answers and reflecting before answering specific questions about evidence supported practices, should all be added to the evaluation process. General questions about TA quality should be avoided and anonymous online evaluations of teaching should not be used for hiring and promotion purposes. Pervasive biases that stem from students' conscious or unconscious social expectations associated with gender will not be completely removed from evaluation of teaching, regardless of our work to apply these interventions. Research on intervention success is much needed if we intend to continue to use and thus hope to improve these tools.

Acknowledgements

The authors would like to acknowledge Erica Mudrak of the Cornell Statistical Consulting Unit (CSCU) for generous support, guidance, and collaboration on statistical analysis of the data and for providing feedback on questions of interpretation and reporting in the manuscript. Two anonymous reviewers also provided important feedback with which we improved the final version of the work shared here. We are grateful to Tom Loiacono at Cornell Engineering IT was instrumental in providing additional data (gender IDs) so that we could examine these important questions. We are also grateful for the work of all the TAs in the Cornell College of Engineering, and support for TA Development from the Cornell College of Engineering.

Author contributions

CE was primarily responsible for designing and framing the study, analyzing the data, and drafting the manuscript. In addition, CE was involved in final check on narrative data coding. DY was responsible for the random selection of the subset of narrative data. DY and KA did the initial coding of those data into themes. KA also conducted a literature search to provide additional citations supporting the assertions made in the paper. DY and KA provided text for the narrative analysis section of the methods and edits to improve the entire manuscript. LS, Director of Engineering Learning Initiatives, supported the development of the work and provided critical insights, suggestions and edits to the developing manuscript throughout the process. All authors read and approved the final manuscript.

Funding

Data collected and analyzed for this study were collected in the regular course of the work we do at Engineering Learning Initiatives in the College of Engineering to improve Teaching Assistant (TA) development. The publication cost was generously covered by the Cornell Open Access Publishing (COAP) Fund of the Cornell University Library.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹Engineering Learning Initiatives, Cornell University, Ithaca, NY, USA. ²Civil and Environmental Engineering, Cornell University, Ithaca, NY, USA. ³Department of Engineering Education, University at Buffalo, Buffalo, NY, USA.

Received: 14 August 2023 Accepted: 19 December 2023

Published online: 29 January 2024

References

- Adams, S., Bekker, S., Fan, Y., Gordon, T., Shepherd, L. J., Slavich, E., & Waters, D. (2022). Gender bias in student evaluations of teaching: 'punish[ing] those who fail to do their gender right.' *Higher Education*, 83(4), 787–807. <https://doi.org/10.1007/s10734-021-00704-9>
- Akin, V., Santillan, S.T., Valentino, L. (2022). Strengthening the STEM Pipeline for Women: An Interdisciplinary Model for Improving Math Identity. *Problems, Resources, and Issues in Mathematics Undergraduate Studies*. <https://doi.org/10.1080/10511970.2022.2032506>
- Almukhambetova, A., Torrano, D. H., & Nam, A. (2023). Fixing the leaky pipeline for talented women in STEM. *International Journal of Science and Mathematics Education*, 21(1), 305–324. <https://doi.org/10.1007/s10763-021-10239-1>
- American Society for Engineering Education. (2022). *Profiles of Engineering and Engineering Technology, 2021*. Washington, DC. <https://ira.asee.org/wp-content/uploads/2022/11/Engineering-and-Engineering-Technology-by-the-Numbers-2021.pdf>
- Aragón, O. R., Pietri, E. S., & Powell, B. A. (2023). Gender bias in teaching evaluations: The causal role of department gender composition. *Proceedings of the National Academy of Sciences*, 120(4), e2118466120. <https://doi.org/10.1073/pnas.2118466120>
- Armstrong, M. A., Jovanovic, J. (2015). Starting at the crossroads: Intersectional approaches to institutionally supporting underrepresented minority women STEM faculty. *Journal of Women and Minorities in Science and Engineering*, 21(2), 141–157. <https://doi.org/10.1615/JWOMINORSCE.NENG.2015011275>
- Baldwin, T., & Blattner, N. (2003). Guarding against potential bias in student evaluations: What every faculty member needs to know. *College Teaching*, 51(1), 27–32. <https://doi.org/10.1080/87567550309596407>
- Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education*, 18(2), 91–106. <https://doi.org/10.1007/s11092-006-9001-8>
- Beischel, W. J., Schudson, Z. C., Hoskin, R. A., & van Anders, S. M. (2023). The gender/sex 3x3: Measuring and categorizing gender/sex beyond binaries. *Psychology of Sexual Orientation and Gender Diversity*, 10(3), 355–372. <https://doi.org/10.1037/sgd0000558>
- Blackburn, H. (2017). The status of women in STEM in higher education: A review of the literature 2007–2017. *Science & Technology Libraries*, 36(3), 235–273. <https://doi.org/10.1080/0194262X.2017.1371658>
- Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. *ScienceOpen Research*. <https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1>
- Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145, 27–41. <https://doi.org/10.1016/j.jpubeco.2016.11.006>
- Buser, W., Batz-Barbarich, C. L., & Hayter, J. K. (2022). Evaluation of women in economics: evidence of gender bias following behavioral role violations. *Sex Roles*, 86(11), 695–710. <https://doi.org/10.1007/s11199-022-01299-w>
- Ceci, S. J., Kahn, S., & Williams, W. M. (2023). Exploring Gender Bias in Six Key Domains of Academic Science: An Adversarial Collaboration. *Psychological Science in the Public Interest*, 24(1), 15–73. <https://doi.org/10.1177/15291006231163179>
- Chatman, J. A., Sharps, D., Mishra, S., Kray, L. J., & North, M. S. (2022). Agentic but not warm: Age-gender interactions and the consequences of stereotype

- incongruity perceptions for middle-aged professional women. *Organizational Behavior and Human Decision Processes*, 173, 104190. <https://doi.org/10.1016/j.obhdp.2022.104190>
- Christensen, R. H. B. (2022). *ordinal—Regression Models for Ordinal Data*. R Package Version 2022.11–16. [Computer software]. Retrieved from <https://CRAN.R-project.org/package=ordinal>
- Clark, S. L., Dyar, C., Inman, E. M., Maung, N., & London, B. (2021). Women's career confidence in a fixed, sexist STEM environment. *International Journal of STEM Education*, 8(1), 56. <https://doi.org/10.1186/s40594-021-00313-z>
- Connell, G. L., Donovan, D. A., & Chambers, T. G. (2016). Increasing the Use of Student-Centered Pedagogies from Moderate to High Improves Student Learning and Attitudes about Biology. *CBE Life Sciences Education*, 15(1), ar3. <https://doi.org/10.1187/cbe.15-03-0062>
- Creswell, J. W., & Creswell, J. D. (2022). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (6th ed.). SAGE Publications, Inc. Retrieved from <https://us.sagepub.com/en-us/nam/research-design/book270550>
- Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3), 573–598. <https://doi.org/10.1037/0033-295X.109.3.573>
- Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS ONE*, 14(2), e0209749. <https://doi.org/10.1371/journal.pone.0209749>
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafo, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. <https://doi.org/10.1073/pnas.1319030111>
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1), 4–27. <https://doi.org/10.1037/0033-295X.102.1.4>
- Grieco, E. G., & Deitz, S. (2023). *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2023* (Special Report NSF No. 23–315). National Center for Science and Engineering Statistics, National Science Foundation. Retrieved from <https://www.nsf.gov/statistics/wmpd>
- Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology*, 29(4), 410–425. <https://doi.org/10.1016/j.cedpsych.2003.11.001>
- Hamrick, K., Falkenheim, J., Hale, K., & Chang, W. (2019). *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2019* (Special Report NSF No. 19–304). National Center for Science and Engineering Statistics, National Science Foundation. Retrieved from <https://ncses.nsf.gov/pubs/nsf19304/digest>
- Hill, C., Corbett, C., & St. Rose, A. (2010). *Why So Few? Women in Science, Technology, Engineering, and Mathematics*. American Association of University Women. <https://eric.ed.gov/?id=ED509653>
- Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117(9), 4609–4616. <https://doi.org/10.1073/pnas.1914221117>
- Hughes, C. C., Schilt, K., Gorman, B. K., & Bratter, J. L. (2017). Framing the faculty gender gap: A view from STEM doctoral students. *Gender, Work & Organization*, 24(4), 398–416. <https://doi.org/10.1111/gwao.12174>
- Hunt, P. K., Dong, M., & Miller, C. M. (2021). A multi-year science research or engineering experience in high school gives women confidence to continue in the STEM pipeline or seek advancement in other fields: A 20-year longitudinal study. *PLoS ONE*, 16(11), e0258717. <https://doi.org/10.1371/journal.pone.0258717>
- Jesse, J. K. (2006). Redesigning science: Recent scholarship on cultural change, gender, and diversity. *BioScience*, 56(10), 831–838. [https://doi.org/10.1641/0006-3568\(2006\)56\[831:RSROC\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2006)56[831:RSROC]2.0.CO;2)
- Khazan, E., Borden, J., Johnson, S., & Greenhaw, L. (2019). Examining Gender Bias in Student Evaluations of Teaching for Graduate Teaching Assistants. *NACTA Journal*, 64(2), 422–427. <https://www.jstor.org/stable/27157815>
- Lattuca, L. R., & Domagal-Goldman, J. M. (2007). Using qualitative methods to assess teaching effectiveness. *New Directions for Institutional Research*, 2007(136), 81–93. <https://doi.org/10.1002/ir.233>
- Llorens, A., Tzovara, A., Bellier, L., Bhaya-Grossman, I., Bidet-Caulet, A., Chang, W. K., Cross, Z. R., Dominguez-Faus, R., Flinker, A., Fonken, Y., Gorenstein, M. A., Holdgraf, C., Hoy, C. W., Ivanova, M. V., Jimenez, R. T., Jun, S., Kam, J. W. Y., Kidd, C., Marcelle, E., ... Dronkers, N. F. (2021). Gender bias in academia: A lifetime problem that needs solutions. *Neuron*, 109(13), 2047–2074. <https://doi.org/10.1016/j.neuron.2021.06.002>
- MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4), 291–303. <https://doi.org/10.1007/s10755-014-9313-4>
- Marks, R. B. (2000). Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education*, 22(2), 108–119. <https://doi.org/10.1177/0273475300222005>
- Mastekaasa, A., & Smeby, J.-C. (2008). Educational choice and persistence in male- and female-dominated fields. *Higher Education*, 55(2), 189–202. <https://doi.org/10.1007/s10734-006-9042-4>
- Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, 17(2), 535–566. <https://doi.org/10.1093/jeaa/jyx057>
- Midway, S., Robertson, M., Flinn, S., & Kaller, M. (2020). Comparing multiple comparisons: Practical guidance for choosing the best multiple comparisons test. *PeerJ*, 8, e10387. <https://doi.org/10.7717/peerj.10387>
- Mitchell, K. M. W., & Martin, J. (2021). Gender Bias in Student Evaluations – Corrigendum. *PS: Political Science & Politics*, 54(1), 192–192. <https://doi.org/10.1017/S1049096520000566>
- Modi, K., Schoenberg, J., & Salmond, K. (2012). Generation STEM: what girls say about science, technology, engineering, and math. Girl Scouts of the USA. Retrieved from https://www.girlscouts.org/content/dam/girlscouts-gsusa/forms-and-documents/about-girl-scouts/research/generation_stem_full_report.pdf
- Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41), 16474–16479. <https://doi.org/10.1073/pnas.1211286109>
- National Center for Education Statistics. (2016). *Integrated Postsecondary Education Data System (IPEDS), Fall 2015 and Fall 2016, Completions component. Indicator 26: STEM degrees* (Archived Survey Materials). U.S. Department of Education. Retrieved from <https://nces.ed.gov/ipeds/use-the-data/annual-survey-forms-packages-archived/2015>
- Peterson, D.A.M., Biederman, L.A., Andersen, D., Ditonto, T.M., Roe, K. (2019) Mitigating gender bias in student evaluations of teaching. *PLoS ONE*, 14(5): e0216241. <https://doi.org/10.1371/journal.pone.0216241>
- Piatek-Jimenez, K., Cribbs, J., & Gill, N. (2018). College students' perceptions of gender stereotypes: Making connections to the underrepresentation of women in STEM fields. *International Journal of Science Education*, 40(12), 1432–1454. <https://doi.org/10.1080/09500693.2018.1482027>
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org/>
- Ritter, B. A., & Yoder, J. D. (2004). Gender differences in leader emergence persist even for dominant women: An updated confirmation of role congruity theory. *Psychology of Women Quarterly*, 28(3), 187–193. <https://doi.org/10.1111/j.1471-6402.2004.00135.x>
- Salmon, U. (2022). Strategies to address gendered racism in science research careers: A scoping review. *Journal for STEM Education Research*, 5(3), 344–379. <https://doi.org/10.1007/s41979-022-00079-1>
- Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014). A campus-wide study of STEM courses: new perspectives on teaching practices and perceptions. *CBE Life Sciences Education*, 13(4), 624–635. <https://doi.org/10.1187/cbe.14-06-0108>
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598–642. <https://doi.org/10.3102/0034654313496870>
- Sprague, J., & Massoni, K. (2005). Student evaluations and gendered expectations: what we can't count can hurt us. *Sex Roles*, 53(11), 779–793. <https://doi.org/10.1007/s11199-005-8292-4>
- Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology*, 42(4), 276–294. <https://doi.org/10.1080/01973533.2020.1756817>
- Syed, M., & Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood*, 3(6), 375–387. <https://doi.org/10.1177/2167696815587648>

- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05." *The American Statistician*, 73(sup1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>
- Weeden, K. A., Gelbgiser, D., & Morgan, S. L. (2020). Pipeline dreams: Occupational plans and gender differences in STEM major persistence and completion. *Sociology of Education*, 93(4), 297–314. <https://doi.org/10.1177/0038040720928484>
- Weisshaar, K. (2017). Publish and Perish? An assessment of gender gaps in promotion to tenure in academia. *Social Forces*, 96(2), 529–560. <https://doi.org/10.1093/sf/sox052>
- Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet*, 393(10171), 531–540. [https://doi.org/10.1016/S0140-6736\(18\)32611-4](https://doi.org/10.1016/S0140-6736(18)32611-4)
- Wright, S. P. (1992). Adjusted P-values for simultaneous inference. *Biometrics*, 48(4), 1005–1013. <https://doi.org/10.2307/2532694>
- Young, S., Rush, L., & Shaw, D. (2009). Evaluating gender bias in ratings of university instructors' teaching effectiveness. *International Journal for the Scholarship of Teaching and Learning*. <https://doi.org/10.20429/ijstl.2009.030219>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.