Sex Roles (2022) 86:695–710
https://doi.org/10.1007/s11199-022-01299-w

ORIGINAL ARTICLE

Evaluation of Women in Economics: Evidence of Gender Bias Following Behavioral Role Violations

Whitney Buser[1] · **Cassondra L. Batz-Barbarich**[2] · **Jill Kearns Hayter**[3]

Accepted: 16 May 2022 / Published online: 15 June 2022

Abstract

Drawing on social role theory (Eagly & Wood, 2016), this paper seeks to understand the nature and causes of gender bias in student evaluations of teaching (SETs) by looking at student evaluations of faculty at two time periods: on the second day of class and on the day after the first exam grade is returned. We seek to understand whether bias exists at the onset of the semester and whether backlash after grading exacerbates any differences. We hypothesized that students would perceive grade feedback more harshly from a female faculty member than a male faculty member due to role congruency expectations of communality in women. The results indicate limited evidence for gender bias at the onset of the semester (the second day of class) and strong evidence for bias against female faculty after the first exam grade is received. This work advances our understanding of when bias develops within the semester and why it may occur. The findings of this study should be of interest to administrators and human resource personnel by ultimately aiding their ability to better manage gender bias in performance evaluations.

Keywords Sex role attitudes · Implicit bias · Gender bias · Higher education · Teaching evaluations · Backlash · Gender role incongruity

Education has long been viewed as a traditionally female-dominated industry, yet higher education remains an exception to this rule (U.S. Department of Labor, Bureau of Labor Statistics, 2020). While women have earned more than half the doctorate degrees conferred for well over a decade (Perry, 2019; U.S. Department of Education, National Center for Education Statistics, 2019; U.S. Department of Education, National Center for Education Statistics, 2020), women continue to be vastly underrepresented in many faculty positions. This is particularly true among tenure-track and tenured positions with only 36% of full professors being women (American Association of University Women [AAUW], 2020). Further, women experience greater underrepresentation within certain fields – such as quantitatively oriented fields like economics, where only 17.5% of full professors are women despite women accounting for 34% of economics bachelor's degrees and 35% of economics graduate degrees (Chevalier, 2019, 2020).

To close this gap, we must better understand *why* the gap persists despite women's educational attainment steadily increasing, and in some fields, even surpassing men (Perry, 2019; U.S. Department of Education, National Center for Education Statistics, 2019, 2020). A myriad of reasons has been put forward to explain why women lag in educational position and prestige (American Association of University Women [AAUW], 2020) while leading in educational attainment in many fields (Perry, 2019; U.S. Department of Education, National Center for Education Statistics, 2019, 2020). These reasons include structural obstacles, lifestyle choices, institutional mindsets, and even individual beliefs

✉ Whitney Buser
buser@gatech.edu

Cassondra L. Batz-Barbarich
cbatz-barbarich@lakeforest.edu

Jill Kearns Hayter
hayter@etsu.edu

[1] School of Economic, Ivan Allen College of Liberal Arts, Georgia Institute of Technology, 221 Bobby Dodd Way NW, Atlanta, GA 30332, USA

[2] Department of Economics, Business, and Finance, Lake Forest College, 555 N Sheridan Road, Lake Forest, IL 60045, USA

[3] Department of Economics and Finance, East Tennessee State University, 325 Sam Wilson Hall, Johnson City, TN 37614, USA

(Barsh & Yee, 2012). Among the various factors considered, student evaluations of teaching (SETs) continue to receive substantial attention among scholars and higher education administrators. Despite growing evidence that a gender bias favoring men is present, SETs continue to be widely used as a key factor in decisions involving hiring, tenure, and promotion (Chávez & Mitchell, 2020), and a substantial contributor to the disparity between male and female faculty's presence and advancement in higher education (Weisshaar, 2017).

This paper uses a quasi-experimental approach to better understand the dynamic nature of gender bias in SETs. Previous work has almost exclusively focused on the end of the semester reviews when official SETs are conducted by universities. By evaluating SETs at only this point in time, the literature has failed to consider when this bias emerges and how it evolves. To address the potentially changing presence and magnitude of gender bias in evaluations, we gather SET data at two points in the semester: the second day of the semester and the day after the first exam grade is returned.

Experimental and quasi-experimental studies have provided some evidence that these gender differences found in SETs do not reflect differences in teaching effectiveness, but instead reflect bias against female faculty (Chávez & Mitchell, 2020). Yet, there has been relatively little work in the literature that has examined what drives this bias. Based on social role theory, this paper seeks to better understand the driving force, or forces, of the gender bias found in SETs by examining evaluations for economics faculty, who have been found to experience exceptionally high levels of gender bias (Felkey & Batz-Barbarich, 2021).

Gender Gaps in Higher Education

Within higher education women are overrepresented in the lowest ranks and among the most unstable faculty positions, including non-tenure track roles and adjunct positions (AAUW, 2020; August & Waltman, 2004; Equal Rights Advocates [ERA], 2003). Women are underrepresented among the most elite, secure, and sought-after roles within higher education: tenure-track positions (AAUW, 2020). Furthermore, women in tenure-track positions are less likely to receive tenure and less likely to be promoted to full professor (Weisshaar, 2017). This is referred to as the academic "leaky pipeline," where women's representation continues to decline the further they advance in their career (Gasser & Shaffer, 2014; Goulden et al., 2011; Mason & Goulden, 2004; Wang & Degol, 2013; Winkler, 2000; Wolfinger et al., 2008).

As challenging as the situation is for women in higher education, gender gaps are exacerbated for women in specific fields. While gender equality has improved in some STEM fields, economics has lagged (Cheryan, et al., 2017). Economics remains an exceptionally challenging field for women to thrive within, and yet it receives substantially less attention in the literature, perhaps because it is viewed as distinct from other STEM fields and instead grouped with social sciences where women are better off comparatively (Boring, 2017; Felkey & Batz-Barbarich, 2021; Ginther & Kahn, 2004; Hale & Regev, 2014; McDowell et al., 2001; Ridgeway & Correll, 2004; Riegle-Crumb & Humphries, 2012). In fact, while other fields have been making strides in women's advancement, economics has remained stagnant over time with employment gains for female faculty at top PhD granting institutions being as small as 2% over ten years (9.7% in 1997 and 11.9% in 2007; Hale & Regev, 2014) and as small as 5% over 20 years across all PhD granting institutions (Chevalier, 2020). Collectively, we focus on economics in this study because of the lack of research in the field and the large gender differences that exist within it.

Gender Gaps in Higher Education Through the Lens of Social Role Theory

Social Role Theory (SRT) provides a useful framework to understand the complexity of gender gaps in the workplace, particularly the inequities in experiences, expectations, and outcomes felt by women at work (Eagly & Wood, 2016). According to SRT, gender inequities are driven by cultural beliefs and expectations for women and men that arise from their distribution into social roles based on physical sex differences (e.g., women as caretakers, men as providers). The overrepresentation of women in low status caregiving social roles leads society to largely hold the belief that women possess the necessary qualities for these roles, such as being friendly, helpful, sensitive, concerned with others, kind, and caring (i.e., more communal qualities). However, the overrepresentation of men in high-status, provider roles leads society to view them as competent, ambitious, assertive, authoritative, and dominant (i.e., more agentic qualities; Eagly & Karau, 2002; Eagly & Wood, 2016). Not only are these qualities used to then describe men and women respectively, but they are also the qualities that men and women are *expected* to have by society.

Relatedly, Role Congruity Theory (RCT) considers the consequences of failing to fulfill these expectations either through the behavior one enacts or the roles they fill. The experience of gender bias is particularly prevalent for women that violate cultural expectations for women's roles and behavior (Eagly & Karau, 2002; Rudman & Phelan, 2008). For example, women in traditionally male-dominated positions (e.g., college professors) or male-dominated fields (e.g., economics) and women that behave in traditionally more agentic ways (e.g., assertive, powerful), are more likely

to experience bias in the form of social backlash (Rudman, 1998). This backlash makes advancement for women more difficult, especially for those in male-dominated fields (Eagly & Karau, 2002; Rudman & Phelan, 2008). Krefting (2003) discussed this conundrum for female academics in which they must be perceived to be competent and authoritative (i.e., powerful, in control) to fulfill their work role in the academic environment but do so at the expense of being perceived as caring and warm (Chávez & Mitchell, 2020; Eagly & Mladinic, 1994; Fiske et al., 2002; MacNell et al., 2015; Williams & Tiedens, 2016).

While RCT suggests there are consequences for violating gender role expectations, Rudman et al. (2012) propose that certain violations may drive women's experience of backlash more so than others. In particular, gender role violations that challenge the traditional gender hierarchy (i.e., men in high-status and women in low-status roles) are less tolerated and therefore produce more severe backlash. They predict that this will be particularly true not only when the *role* violates status expectations, but when their *behaviors* within that role do as well. These propositions are summarized by the Status Incongruity Hypothesis (SIH) which suggest the presence of backlash will be greatest when women are in high-status roles and display dominance (Rudman et al., 2012). That is, status incongruity exacerbates existing backlash from role and behavior violations.

Gender Backlash in SETs

Weisshaar (2017) found that while productivity differences account for a portion of the gender difference in tenure decisions, it does not account for all of it. Instead, there is evidence that gendered processes, such as women's work being devalued or scrutinized more harshly, significantly explained differences in promotion and tenure decisions. This devaluation and greater scrutinization may be evidence of backlash against high-status women. This possibility is supported by a myriad of experimental studies that manipulate the gender of the faculty member and find devaluations of women as compared to men when evaluating identical teaching, research, or service records (Knobloch-Westerwick et al., 2013; Steinpreis et al., 1999).

Historically researchers made claims that SETs are valid, reliable, and unaffected by potential instances of bias and backlash (Aleamoni, 1999; Cohen, 1981; Marsh, 1987; Wilson et al., 1997). However, more recent work has challenged this assertion. Learning outcomes (e.g., grades, hours studied) and other objective indicators of effectiveness in teaching do not vary depending on gender of the instructor, yet female faculty systematically receive lower teaching evaluations than their male peers (Boring, 2017; Chávez & Mitchell, 2020; Mengel et al., 2019). Research has also shown that women may invest more time in teaching than men, which challenges this assertion as well (Misra et al., 2010; Winslow, 2010).

Previous work has supported bias and backlash as a driving force underlying these differences in SETs. A study conducted by MacNell et al. (2015) found significant gender bias in SETs for an online course where students believed the course was taught by a male faculty member or a female faculty member, while in reality the course was taught by the same faculty. Their findings suggest that bias persists when one controls for teaching quality. Arbuckle and Williams (2003) also found evidence of gender differences in SETs when students were exposed to a neutral stick figure and voice accompanied with a bio that manipulated the instructor's gender. Interestingly, gender bias is not confined to evaluating the instructor alone. Research by Mengel et al. (2019) found that teaching materials, such as textbooks, that are identical across classes are rated more negatively in courses led by female faculty as compared to those led by male faculty.

This bias in SETs is larger for more quantitatively heavy courses (e.g., mathematical courses; Mengel et al., 2019) and courses that are predominantly filled with male students who tend to have a greater bias towards female faculty. In fact, male students are 30% more likely to rate their male faculty as excellent compared to their female faculty (Abel, 2019; Mengel et al., 2019). McPherson et al. (2009) found gender as a significant determinant of student evaluation ratings in both upper-level and lower-level economics courses. Their results showed male professors received higher student evaluation ratings compared to female professors in all levels of economics courses. Felkey and Batz-Barbarich (2021) examined this question meta-analytically, seeking to compare economics (a male-dominated and quantitatively heavy subject) to its peer social sciences that are less dominated by men and substantially less quantitative in nature. They found significant gender bias exists in SETs favoring men in economics, but not in other social sciences. The previous discussion further supports the notion that while gender bias exists on a broad level within SETs, it appears to be more severe in particular fields, including economics.

The Present Study

Though it is well-established that gender bias influences SETs, several questions remain unanswered. Using a short-term longitudinal approach, we seek to better deduce whether bias, and subsequent backlash, exists at the onset of the semester presumably due to role-based incongruity violations or whether backlash is in response to behavior-based incongruity. Evidence of bias on the second day of the course, when students have only observed a small number of

behavioral cues, would suggest the presence of role-based or status-based gender bias. Evidence of gender bias following female faculty's assertion of power and dominance via giving feedback suggests behavior-based gender bias, or rather backlash based on women *behaving* in non-traditional ways. Further understanding of the driving forces behind gender bias in SETs will better equip us to meaningfully explain, minimize, and perhaps even eradicate bias in SETs.

Study Hypotheses

Gender Differences at Time 1

Sex role expectations lead to perceptions that women are, and should be, more communal (e.g., caring, supportive) and men are, and should be, more agentic (e.g., assertive, powerful) and should occupy roles that align with these qualities (Eagly & Karau, 2002; Hentschel et al., 2019). As such, one might anticipate that evaluations on the second day of class (Time 1) will differ as a function of instructor gender in ways that align with these expectations. This is due to limited additional information to base assessments and little time to witness behaviors counter to these expectations at this point in the semester. However, previous work has found that backlash is driven not only by behavior, but also role incongruity which is exacerbated by status incongruity (Rudman, et al., 2012). We anticipate that women in a faculty role (i.e., a position of status and power)– particularly in a male-dominated field– would experience backlash simply for filling a role that violates expectations of the type of roles women should occupy. We suspect that this role-based backlash would be made apparent via significant gender differences in SETs.

Despite stereotypes that women are more communal and men more agentic, we do not anticipate evaluations to reflect this pattern, and instead expect that women will receive lower SET scores across all qualities assessed, which is consistent with the existing SET literature (Boring, 2017; Chávez & Mitchell, 2020; Mengel et al., 2019). More specifically, and consistent with SRT, we anticipate women will be punished for violating gender-role expectations by being in a gender incongruent role. We predict this will result in lower scores on communal qualities than men. Second, despite women being in a role that aligns more with agency, their identity as a woman will still likely lower their rating of this attribute as well. Whereas they may have been rated as more agentic than women in other roles, we anticipate they will still be rated as less agentic than men because less agency is attributed to women in general. In other words, we anticipate a "catch-22" situation for women, where their role is too agentic to be viewed as communal, but their gender is too communal to be rated as more agentic. Lastly, for gender-neutral attributes that reflect a general impression of faculty and their courses, we anticipate again that the backlash due to perceived gender role-incongruency would produce lower ratings for women compared to men on these attributes as well.

As such, we predict:

> ***H1:*** Female faculty will be rated lower on (a) gender-neutral qualities (i.e., *Recommend Course*, *Recommend Instructor*, and *Interesting*), (b) agentic qualities (i.e., *Knowledgeable*, *Challenging*), and (c) communal qualities (i.e., *Caring*; *Approachable*) than male faculty at Time 1, after controlling for other potential explanatory factors.

Gender Differences at Time 2 and Across Time

Based on the SIT framework, Rudman et al. (2012) predict backlash is likely to be more severe when behaviors within the role also violate status expectations. As such, women in a high-status role such as a faculty member who also enact a behavior that violates their gender roles will likely experience increased backlash. This is particularly true when the behavior enacted highlights their power and dominance in the classroom (i.e., providing feedback; Rudman et al., 2012). While we note that providing feedback is a task expected from a faculty member, we propose that this behavior would be less palatable to students depending upon the gender of the faculty member. This prediction is based on past research that found when fictitious female managers, compared to male managers, delivered identical critical feedback, participants perceived that feedback to be significantly less accurate and less appropriate (Abel, 2019). Abel explains this based on the gendered expectations of the participants – they were three times more likely to associate positive feedback with female managers and twice as likely to associate critical feedback with male managers. Therefore, female managers that provide critical feedback violate gender expectations, and face consequences for doing so. Related work by Sinclair and Kunda (2000) found that women are rated as less competent than men in SETs after providing negative feedback to students, but not after providing positive feedback. As such, we anticipate something similar to happen for female faculty, such that gender differences in communal, agentic, and gender-neutral qualities will not only remain at Time 2 (H2), but that these differences will be exacerbated with women being rated significantly lower on these qualities at Time 2 as compared to Time 1 (H3).

Importantly, we want to note that there is no empirical or theoretical rationale for anticipating differences in men's ratings over time. As such we base our hypothesis solely on our expectations for women relative to men. In other words, even if men's ratings did decrease – which the literature gives us no reason to anticipate – we would still

predict that women's decrease would be more severe due to the theories highlighted above and past empirical work. We suspect that this behavior-based backlash would be made apparent via significant gender differences in SETs at Time 2, differences that we anticipate would be greater than at Time 1.

H2: Female faculty will be rated lower than male faculty on (a) gender neutral, (b) agentic, and (c) communal qualities at Time 2, after controlling for other potential explanatory factors.

H3: Differences in mean evaluation scores between men and women will be larger at Time 2 compared to Time 1 on (a) gender-neutral, (b) agentic, and (c) communal qualities, after controlling for other potential explanatory factors.

Method

Participants

Participants were undergraduate students ($N = 1{,}191$; 696 men; 494 women; 81% White) enrolled in introductory-level economics courses taught by seven different faculty members (men = 3; women = 4) at five unique institutions over three semesters. The institutions included one state university, one large regional university, and three private liberal arts colleges, all of which are in the United States. At both the regional university and one liberal arts college, we have data from both one male and one female instructor. Of the four female faculty, none were underrepresented minorities (URM). Of the three male faculty, one was an URM (i.e., Black).

All participants were enrolled in an introductory course on the principles of economics. Despite the teaching evaluation literature clearly showing a difference in evaluation responses across course type, level, and subject (Liaw & Goh, 2003; Macfadyen et al., 2016), to our knowledge, this the first study which examines data taken from the same course over different faculty and institutions. Principles of economics courses vary little in content, structure, and approach. We verified that our faculty all taught almost identical material, at a very similar pace, and assessed with two to three multiple choice mid-semester exams and a multiple-choice final exam. Introduction to economics was chosen due to similar course content, as well as the fact that it is taken as a requirement at most colleges and universities (inclusive of the institutions in our sample) and the field is male dominated (Jonung & Stahlberg, 2009). Table 1 provides an overview of the demographic information for the sample.

Table 1 Demographic Variables: Frequencies and Means

Variables	Frequency/Mean				
	N		*SD*	Min	Max
Full Sample					
Female Instructor	1,191	37.6%			
Female Student	1,188	41.5%			
Mother Education	1,191	44.2%			
First Year	1,186	49.3%			
Time 1	1,191	53.1%			
Time 2	1,191	46.9%			
Expected Grade	1,174	3.537	.62	1	4
Interest Economics	1,183	2.074	.91	0	4
Female Instructor Courses					
Female Student	741	43.7%			
Mother Education	743	49.9%			
First Year	739	56.0%			
Time 1	743	54.1%			
Time 2	743	45.9%			
Expected Grade	730	3.555	.61	1	4
Interest Economics	739	2.058	.93	0	4
Male Instructor Courses					
Female Student	741	43.7%			
Mother Education	743	49.9%			
First Year	739	56.0%			
Time 1	743	54.1%			
Time 2	743	45.9%			
Expected Grade	730	3.555	.61	1	4
Interest Economics	739	2.058	.93	0	4

Measures

Instructor Traits in SETs

Because different institutions use different measures and scales for conducting teaching evaluations, a standard survey was created for this study. Using the same survey also allows for a direct comparison across participants at multiple institutions. The items comprising the survey were informed by a literature review on the different characteristics that have been used to detect gender bias in faculty evaluations (Bachen et al., 1999; Boring, 2017; MacNell et al., 2015). Specifically, students were asked to evaluate their course instructor on seven items using a 5-point scale from 0 (*Strongly Disagree*) to 4 (*Strongly Agree*). Specifically, 0 represents "*Strongly Disagree*," 1 represents "*Disagree*," 2 indicates "*Neutral*," 3 signifies "*Agree*," and 4 represents "*Strongly Agree*." Each item was analyzed separately. Three items assess gender neutral content relevant to teaching: *Recommend Course* (i.e., "I would recommend this course to other students looking for a worthwhile

course;" *Recommend Instructor* (i.e., "I would recommend this instructor to other students looking for a good teacher);" and *Interesting* ("My instructor is interesting"). Two items assess gendered traits related to agency (Bachen et al., 1999; Boring, 2017): *Knowledgeable* ("My instructor is knowledgeable") and *Challenging* ("My instructor challenged me intellectually"). Two items assess gendered traits related to communality (Bachen et al., 1999; MacNell et al., 2015): *Approachable* ("My instructor is approachable") and *Caring* ("My instructor is caring").

Covariates

All covariates were selected based on previous research on outcomes relevant for Introduction to Economics classes (i.e., Al-Bahrani et al., 2020; Rousu et al., 2015) and appear exclusively as controls in the regression analysis. Covariates included instructor gender, participant gender, whether the participant was non-white, the education level of the participant's mother, whether the participant was in their first year of college, and participants' interest in economics. In addition, participants were asked to report their expected grade in the course as either an A, B, C, D, or F. Due to IRB constraints, matching individual responses with grades proved impossible and we were not allowed to collect this data. Instead, we collected two key pieces of grade information: participants' expected course grade and instructor level mean exam score based on the exam given immediately prior to Time 2 evaluations – both of which we controlled for in the study. Despite not being able to collect individual grade data for participants, we argue that controlling for expected final course grade is more predictive of course evaluation results than actual grade received on one exam. That is, some students are more accurate than others in predicting their final grade based on one exam score. Overall, participants expected to earn high grades, but expectations did adjust downward after the first exam.

Many other factors were taken into consideration but do not appear in the control vectors of our analysis for various reasons. Students were asked if their economics course was a required course for their major or minor in addition to whether it was taken as a requirement at all. It was clear from the responses (with most students answering that economics was a major requirement) that students were answering in the affirmative even if they were taking economics as part of their core requirements. This factor was therefore dropped from the analysis. The analysis also dropped information regarding SAT and ACT scores due to low response rates. Due to low response rates, we also dropped information regarding father's education level, preparation hours for the course, number of absences, time spent working outside of class at a job, and specific major.

Procedure

Data were collected from the participants using two paper surveys at two time points: Survey 1 was administered on the second day of class (Time 1) to examine participants' early impressions of their faculty. Survey 2 was administered the day after receiving the first exam grade (Time 2) to examine whether their impression of the faculty had changed after receiving feedback from them.

Following the distribution of a consent form, a survey administrator informed participants that their responses would be analyzed at a separate institution and their participation was voluntary. Participants were offered 2 bonus points for either participating in the study or participating in an alternate activity. No participants chose to opt out of the study. Our sampling was also unique in that all participants were given their surveys in written form and allowed class time to complete their surveys thus eliminating complications from online administration. The study was approved by the appropriate Institutional Review Board and all participants signed informed consent documents.

Instructors were not present when students were asked to participate in the study and members of the research team were provided with a script to standardize the introduction of data collection across participants. Participants were informed that two bonus points on the subsequent exam would be awarded if the survey was completed by more than 80% of the class. Participants were also told that their instructor would not have access to individual survey responses and the data would be anonymized before analysis. If students did not wish to participate, they could simply turn in a blank survey form. The survey administrator collected all surveys (blank or completed) and placed them in a stamped, addressed envelope and immediately sealed and mailed the surveys to the investigator.

It is also important to note that due to variable class attendance, not every student completed both surveys, but roughly one-half of the data comes from each of the surveys. For paired analysis, any data not paired was dropped from our analysis. There was a single standard by which we judged if a statistic was significant or not: the .05 level. There was no adjustment of what was considered significant; all significance testing is uniform.

Analytic Strategy

We examined the role of instructor gender on participants' evaluations of faculty at two time periods across the seven SET items. For each outcome variable, mean evaluation scores were calculated for women at Time 1, women at Time 2, men at Time 1, and men at Time 2. Independent sample *t*-tests were used to identify any significant gender differences in means between men and women at Time 1

and between men and women at Time 2. Then, to test the impact of time and feedback by instructor gender, paired sample t-tests were used comparing male instructor ratings at Time 1 and Time 2 and female instructor ratings at Time 1 and Time 2.

We used regression analysis to check that the results from the unadjusted mean comparisons held when controlling for participant characteristics. Due to the nature of the dependent variable, all regression specifications were run as both an ordered probit and a linear model. Due to the nature of the non-interdependence of data, all models were specified to be multilevel mixed regression grouped by classroom. Because results were robust to either mixed ordered probit or mixed linear, the multilevel linear models are presented for simplicity. Because we use two-time period longitudinal data, a difference in difference (DID) model was also run. To clearly demonstrate instructor gender differences at each point in the semester we stratified the data by time. Table 3 presents results of a multilevel linear model grouped at the classroom level on the second day (Time 1) only. Table 4 presents the post feedback regressions using data from Time 2.

The specification for these regressions is:

$$\begin{aligned} Evaluation_{ijt}{}^{k} = {} & \beta_0 + \beta_1 Female\ Instructor_{ijt} \\ & + \beta_2 Female\ Student_{ijt} \\ & + \beta_3 \underline{X}_{ijt} + \varepsilon_{ijt} \end{aligned}$$

where $Female\ Instructor_{ijt}$ is a dummy variable equal to 1 if the instructor is female, $Female\ Student_{ijt}$ is a dummy variable equal to 1 if the student performing the evaluation is female, $\underline{X}_{ijt}$ is a set of student-specific control variables, and ε_{ijt} is the stochastic error term. Subscripts indicate student i at institution j in time t (Time 1 or 2). The superscript, k, represents the evaluation criterion that is being assessed. Regressions were also specified as a logistic model where the dependent variable was collapsed into a dichotomous variable where 1 indicated a perfect evaluation score and 0 indicates otherwise. Please see Tables S1 and S2 in the online supplement for these results.

The vector of control variables includes participant characteristics such as ethnicity (coded as 1 for non-white and 0 for white), whether the student's mother has a college education or higher (1 if mother's education is college or greater, 0 if mother's education is less than college level), whether the student is a first-year undergraduate (1 if student is first year, 0 otherwise), the expected grade of the student in the course at the time of the survey (coded as A = 4.0, B = 3.0, C = 2.0, D = 1.0, or F = 0), and the level of student interest in economics at the time of the survey coded as (4 = very interested, 3 = somewhat interested, 2 = neutral, 1 = uninterested, 0 = very uninterested).

Initially, we also stratified our sample by instructor gender to isolate and compare the temporal changes for men and women as well as to see the direct effects of student gender on men and women individually. These regressions yielded no interesting results regarding the effect of student gender or temporal effects of female faculty. Male faculty were shown to improve over time, which can be seen in Fig. 1a–g. Therefore, these analyses were not included in the main article. Please see Table S3 and S4 in the online supplement for detailed information on the regressions stratified by student gender.

In addition to multilevel linear regressions grouped by classroom and stratified by time, we also ran a difference in difference (DID) model to see if the gender bias at Time 1 changed significantly compared to Time 2. Results for the DID specification appear in Table 5.

The specification of the DID is:

$$\begin{aligned} Evaluation_{ijt}{}^{k} = {} & \beta_0 + \beta_1 Time\,2_{ijt} + \beta_2 Female\ Instructor_{ijt} \\ & + \beta_3 Time2 * Female\ Instuctor_{ijt} \\ & + \beta_4 Female\ Student_{ijt} + \beta_6 \underline{X}_{ijt} + \varepsilon_{ijt} \end{aligned} \quad (1)$$

where $Time\,2_{ijt}$ is a dummy variable equal to 1 when the evaluation measures the post feedback response, $Time\,2 * Female\ Instuctor_{ijt}$ is the difference and difference interaction term. In specification (1) the coefficient on the interaction term shows the difference in the difference of means between genders and times. That is,

$$\begin{aligned} \beta_3 = {} & \left(\bar{y}Female\ Instructor, Time\,2 - \bar{y}Female\ Instructor, Time\,1\right) \\ & - \left(\bar{y}Male\ Instructor, Time\,2 - \bar{y}Male\ Instructor\ Time\,1\right) \end{aligned}$$

Results

Preliminary Analysis

We compared the characteristics of students by instructor gender to ensure the comparability of our data between male and female instructors. For Time 2 comparisons we sought to explore if there were differences between male and female instructors that may be driving any observed differences in evaluations. We compared instructors' mean scores for the exam that was given prior to the SETs collected at Time 2. There was no significant difference between male ($M = 74.92$, $SD = 3.12$) and female instructors ($M = 75.33$, $SD = 3.38$), $t(466) = .16$, $p = .99$ on the exam grades given. This allows us to rule out the possibility that the difference in evaluations is motivated by female faculty being more stringent graders. Please see Table S5 in the online supplement for a correlation matrix between variables.

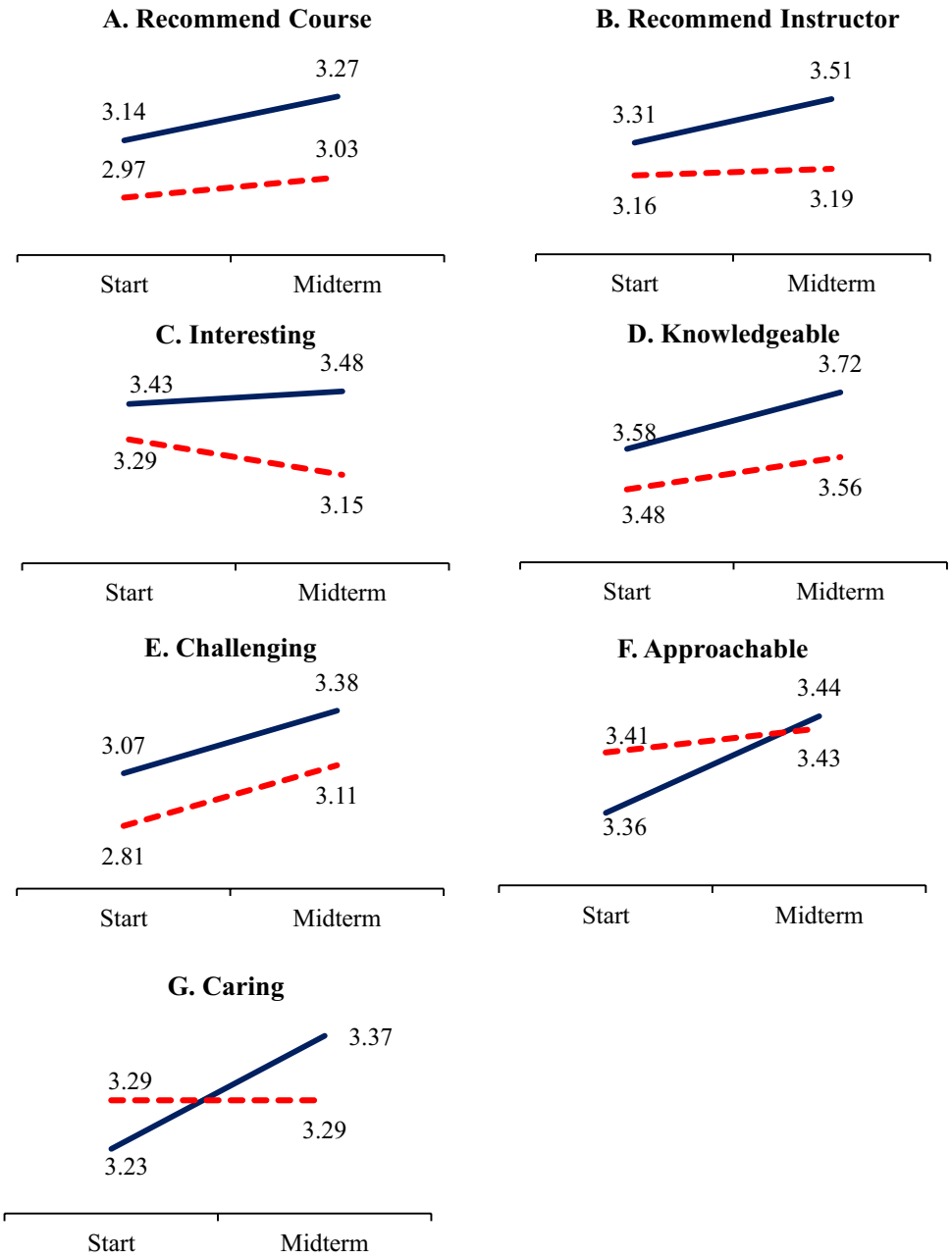

Fig. 1 **A–G** Unadjusted Means for Student Evaluations Responding to Each SET Item by Gender and Time Period

Independent and Paired Sample t-tests

Table 2 presents the results of the independent *t*-tests and paired sample *t*-tests comparing male and female faculty across the study variables over the two time points.

Instructor Gender Differences at Time 1

In support of Hypothesis 1a, an independent sample *t*-test indicated that female instructors were rated significantly lower at Time 1 than male instructors on all three gender-neutral qualities: *Recommend Course*, $t(623) = 2.81$, $p = .005$, *Recommend Instructor*, $t(622) = 2.41$, $p = .016$, and *Interesting*, $t(623) = 2.40$, $p = .017$). In partial support of Hypothesis 1b, a significant difference was observed for one of the agentic qualities. An independent sample *t*-test indicated that female instructors were rated significantly lower than male instructors at Time 1 on *Challenging*, $t(622) = 3.88$, $p < .001$), but no significant difference was observed for *Knowledgeable*, $t(623) = 1.85$, $p = .06$. Failing to support Hypothesis 1c, no differences were observed for the communal qualities. An independent sample *t*-test indicated that female instructors were not rated significantly higher than male instructors at Time 1 on *Caring*, $t(620) = -.94$ $p = .34$), or *Approachable*, $t(622) = -.84$, $p = .39$).

Table 2 Summary of Results for Comparisons by Gender Group and Time on SET Variables

	Men		Women						
Variable	*M*	*SD*	*M*	*SD*	Comparison	*t* ratio	df	*p*	Cohen's *d*
Recommend Course									
Time 1	3.14	.78	2.97	.79	G	2.81	446	.005	.22
Time 2	3.27	.84	3.03	.91	G	3.24	446	.001	.27
					TW	–.78	445	.43	.07
					TM	–2.07	443	.039	.16
Recommend Instructor									
Time 1	3.31	.75	3.16	.76	G	2.41	446	.016	.20
Time 2	3.51	.76	3.19	.89	G	4.53	446	< .001	.39
					TW	–.34	446	.73	.03
					TM	–3.50	446	< .001	.26
Interesting									
Time 1	3.43	.73	3.29	.71	G	2.4	444	.02	.19
Time 2	3.48	.71	3.15	.81	G	5.0	443	< .001	.43
					TW	1.92	446	.06	.18
					TM	–.86	446	.39	.07
Knowledgeable									
Time 1	3.58	.68	3.48	.61	G	1.85	446	.06	.15
Time 2	3.72	.59	3.56	.58	G	3.11	446	.002	.27
					TW	–1.53	445	.12	.11
					TM	–3.04	446	.002	.21
Challenging									
Time 1	3.07	.78	2.81	.78	G	3.88	446	< .001	.33
Time 2	3.38	.67	3.11	.78	G	4.30	445	< .001	.37
					TW	–3.93	443	< .001	.38
					TM	–5.74	445	< .001	.38
Approachable									
Time 1	3.36	.82	3.41	.65	G	–.84	444	.39	.07
Time 2	3.44	.76	3.43	.71	G	.10	446	.92	.01
					TW	–.24	445	.81	.03
					TM	–1.28	446	.21	.03
Caring									
Time 1	3.23	.83	3.29	.70	G	–.94	443	.34	.08
Time 2	3.37	.80	3.29	.74	G	1.08	446	.27	.10
					TW	.03	445	.94	.00
					TM	–2.19	446	.03	.02

t-ratios represent independent samples t-tests for gender group comparisons and paired samples t-tests for time point comparison

SET student evaluations of teaching, *G* comparison by gender group, *TW* comparison by time point for women, *TM* comparison by time point for men

Instructor Gender Differences at Time 2

In support of Hypothesis 2a, an independent sample *t*-test indicated that female instructors were rated significantly lower at Time 2 than male instructors on all three gender-neutral qualities: *Recommend Course*, $t(550) = 3.24$, $p = .001$), *Recommend Instructor*, $t(550) = 4.53$, $p < .001$), and *Interesting*, $t(549) = 5.00$, $p < .001$). In support of Hypothesis 2b, a significant difference was observed for both agentic qualities. An independent sample *t*-test indicated that female instructors were rated significantly lower than male instructors at Time 2 on *Knowledgeable*, $t(551) = 3.11$, $p = .002$), and *Challenging*, $t(549) = 4.30$, $p < .001$). Inconsistent with Hypothesis 2c, no significant differences were observed for the communal qualities. An independent sample *t*-test indicated no differences between female and male instructors at Time 2 for *Caring*, $t(549) = 1.08$, $p = .27$), or *Approachable*, $t(550) = .10$, $p = .92$).

Within Gender Comparisons Between Time 1 and Time 2

To examine Hypotheses 3 a, b, and c, we first used paired sample *t*-tests to examine differences in the ratings between Time 1 and Time 2 for male and female instructors (for further tests see "Regression Analyses" below). No support was found for Hypothesis 3a, apart from one finding for male instructors. For the gender-neutral qualities for female instructors, ratings did not differ between Time 1 and Time 2 for *Recommend Course*, $t(441) = -.78$, $p = .41$, *Recommend Instructor*, $t(441) = -.34$, $p = .73$), or *Interesting*, $t(440) = 1.92$, $p = .06$. For the gender-neutral qualities for male instructors, ratings did not differ between Time 1 and Time 2 for *Recommend Instructor,* $t(731) = -3.50$, $p < .001$), or Interesting, $t(732) = -.86$, $p = .42$); however, a significantly higher rating was observed for *Recommend Course* at Time 2 compared to Time 1, $t(732) = -2.07$, $p = .039$.

Hypothesis 3b was generally supported, apart from one finding for female instructors. For the agentic qualities for female instructors, ratings were higher at Time 2 compared to Time 1 for *Challenging*, $t(440) = -3.98$, $p < .001$, but no difference was observed for *Knowledgeable*, $t(442) = -1.53$, $p = .12$. For the agentic qualities for male instructors, ratings were higher at Time 2 compared to Time 1 on *Knowledgeable,* $t(732) = -3.04$, $p = .002$, and *Challenging*, $t(731) = -5.74$, $p < .001$.

No support was found for Hypothesis 3a, apart from one finding for male instructors. For the communal qualities for female instructors, ratings did not differ for *Approachable*, $t(441) = -.24$, $p = .81$), or *Caring*, $t(438) = .03$, $p = .98$). For the communal qualities for male instructors, ratings did not differ for *Approachable*, $t(731) = -1.28$, $p = .21$; however, a significantly higher rating was observed at Time 2 compared to Time 1 for *Caring*, $t(731) = -2.19$, $p = .03$. Figure 1a–g give provide a visual representation of the patterns for each SET variable.

Regression Analyses

Instructor Gender Differences at Time 1

The regression analysis further examined the hypotheses tested above while controlling for student characteristics. For Hypothesis 1a, 1b, and 1c, we predicted that women would receive lower ratings than men at Time 1 on gender neutral, agentic, and communal qualities. Table 3 presents the results for Time 1 when subjecting the data to multilevel regression at the classroom level. The non-significant effect for the female faculty coefficient indicated no difference in instructor ratings as a function of instructor gender at Time 1 when controlling for participant characteristics and using multilevel modeling to account for interdependence among data.

Instructor Gender Differences at Time 2

For Hypothesis 2a, 2b, and 2c we predicted that female instructors would be rated lower than male instructors at Time 2 on gender neutral, agentic, and communal qualities. As seen in Table 4, the multilevel regressions support Hypothesis 2a and 2b for gender-neutral and agentic qualities of evaluation. These results indicate that women do experience gender bias after feedback when controlling for student characteristics and non-interdependence of the classroom. However, there was no effect of instructor gender for communal qualities (Hypothesis 2c).

Difference in Instructor Gender Differences from Time 1 to Time 2

Table 5 presents the result for the difference in difference approach using observations from a balanced panel of the entire sample. These results continue to support the pattern

Table 3 Time 1 Multilevel Regressions Grouped by Classroom

Variable	Recommend Course	Recommend Instructor	Interesting	Knowledgeable	Challenging	Approachable	Caring
Female Instructor	–.19 (.15)	–.15 (.17)	–.19 (.14)	–.14 (.11)	–.23 (.18)	.03 (.11)	–.01 (.09)
Female Student	.05 (.07)	.03 (.07)	.16 (.06)	.09 (.06)	.11 (.07)	.18 (.07)	.16 (.07)
Non-White Student	.13 (.09)	.07 (.09)	.07 (.08)	.08 (.08)	.08 (.09)	.01 (.09)	.06 (.09)
Mom Education	–.10 (.07)	–.17 (.07)	–.00 (.06)	–.067 (.06)	–.07 (.07)	.03 (.07)	–.02 (–.07)
First Year	–.08 (.08)	.04 (.08)	.00 (.07)	.06 (.06)	–.03 (.08)	–.08 (.07)	–.11 (.07)
Expected Grade	.13 (.07)	.26** (.07)	.20* (.07)	.07 (.06)	.29** (.07)	.19*(.07)	.23*(.07)
Econ Interest	.20***(.04)	.08 (.04)	.07 (.04)	.08 (.03)	.08 (.04)	.06 (.04)	.08 (.04)
Constant	2.28** (.31)	2.24**(.32)	2.53**(.28)	3.12**(.25)	1.81**(.32)	2.52**(.29)	2.25**(.29)
Observations	477	476	477	477	476	477	475
R-Squared	.10	.11	.09	.08	.10	.09	.09

Robust standard errors in parentheses. Number of groups $k = 9$

*$p < .05$; **$p < .01$; ***$p < .001$

Table 4 Time 2 Multilevel Regressions Grouped by Classroom

Variable	Recommend Course	Recommend Instructor	Interesting	Knowledgeable	Challenging	Approachable	Caring
Female Instructor	–.26*** (.08)	–.33*** (.11)	–.38*** (.13)	–.19** (.08)	–.32** (.14)	.05 (.16)	–.10 (.13)
Female Student	.01 (.08)	.08 (.08)	.06 (.07)	.02 (.05)	.01 (.07)	.02 (.07)	.04 (.07)
Non-White Student	.09 (.10)	–.02 (.10)	.08 (.09)	.13 (.07)	–.03 (.09)	.04 (.09)	.06 (.09)
Mom Education	–.04 (.08)	–.13 (.08)	–.12 (.07)	–.09 (.05)	–.04 (.07)	–.05 (.07)	–.07 (.07)
First Year Student	–.07 (.08)	–.05 (.08)	–.03 (.07)	.04 (.06)	–.03 (.07)	.03 (.08)	.02 (.08)
Expected Grade	.28*** (.06)	.26*** (.06)	.02 (.05)	.04 (.04)	–.01 (.05)	.14** (.05)	.21*** (.05)
Economics Interest	.26*** (.05)	.15*** (.04)	.19*** (.04)	.08 (.03)	.08 (.04)	.10 (.04)	.10 (.04)
Constant	1.71*** (.22)	2.37*** (.22)	3.05*** (.21)	3.36*** (.16)	3.37*** (.21)	2.70*** (.23)	2.42*** (.22)
Observations	479	480	478	480	479	479	478
R-Squared	.09	.10	.11	.09	.08	.09	.10

Robust standard errors in parentheses. Number of groups $k = 9$

*$p < .05$; **$p < .01$; ***$p < .001$

of findings indicating that female instructors are rated lower than their male counterparts on all agentic (*Knowledgeable*, *Challenging*) and gender-neutral qualities (*Recommend Course*, *Recommend Instructor*, *Interesting*). The interaction term between female instructor and Time 2 carries the expected negative sign, indicating that there is more gender bias at Time 2 than at Time 1, however the coefficient is not statistically significant. That is, if we compare the difference in SET ratings by gender at Time 1 to the difference in SET ratings at Time 2, we see the differences get larger, but not significantly larger. Based on the coefficient for the interaction terms in the DID model alone, we found no support for Hypothesis 3a, 3b, and 3c.

However, there are results that do support Hypothesis 3a, 3b, and 3c. First, when comparing the coefficients on female instructor regressions in Table 3 (negative, nonsignificant) and Table 4 (negative, significant), we see a change from no significant bias to significant bias. Second, Fig. 1a–g also reveal a widening of differences between men and women for *Recommend Course*, *Recommend Instructor*, *Interesting*, and *Knowledgeable*. More specifically, we see that men increase in their ratings for all characteristics between Time 1 and Time 2 indicating that students see men more favorably as time goes on, which does not happen for women. Further, we see that women are rated as significantly less *Interesting* from Time 1 to Time 2. While this evidence is certainly not as strong, these patterns do align with the basic premise of Hypothesis 3a, 3b, and 3c. However, ultimately, we cannot say that we found support for Hypothesis 3a, 3b, and 3c due to the nonsignificant difference in difference regressions tests.

Exploratory Analyses

A few things are interesting to note about the control variables that appear in each regression table. One of the most consistent findings across the regressions is that Expected

Table 5 Difference in Difference Estimation

Variable	Recommend Course	Recommend Instructor	Interesting	Knowledgeable	Challenging	Approachable	Caring
Time 2	.13 (.07)	.26*** (.06)	.01 (.06)	.13 (.05)	.33*** (.06)	.08 (.07)	.16 (.06)
Female Instructor	–.22** (.07)	–.17 (.07)	–.23*** (.07)	–.16* (.06)	–.30*** (.07)	.04 (.07)	.00 (.07)
Time 2*FemIns	–.05 (.11)	–.18 (.10)	–.14 (.10)	–.03 (.08)	–.00 (.10)	–.04 (.09)	–.12 (.10)
Female Student	.06 (.05)	.09 (.05)	.13** (.05)	.07 (.04)	.07 (.05)	.12 (.05)	.12 (.05)
Non-White Student	.12 (.08)	.04 (.07)	.06 (.07)	.10 (.06)	.01 (.07)	.03 (.06)	.06 (.07)
Mom Education	–.07 (.05)	–.12 (.05)	–.04 (.05)	–.08 (.04)	–.09 (.05)	.03 (.05)	–.02 (.05)
First Year	–.11 (.05)	–.05 (.05)	–.02 (.05)	.04 (.04)	–.04 (.04)	–.06 (.05)	–.06 (.05)
Expected Grade	.26*** (.05)	.29*** (.05)	.07 (.04)	.04 (.04)	.05 (.05)	.19*** (.04)	.24*** (.05)
Econ Interest	.23*** (.03)	.11*** (.03)	.14*** (.03)	.09*** (.02)	.11*** (.03)	.07* (.03)	.08* (.03)
Constant	1.72*** (.19)	2.04*** (.19)	2.86*** (.17)	3.19*** (.15)	2.67*** (.19)	2.47*** (.17)	2.17*** (.19)
Observations	956	956	955	957	955	956	953
R-squared	.13	.10	.09	.06	.10	.04	.06

Robust standard errors in parentheses

*$p < .05$; **$p < .01$; ***$p < .001$

Grade and Interest in Economics are by far the strongest predictors of how a professor will be rated. While our gender differences still hold despite controlling for these two important predictors, it cannot be overlooked that the strongest driver of a professor's rating is a student's perception of non-instructor qualities. This is an important contribution to the literature on teaching evaluations and should be considered by administrators when interpreting evaluation results. Also, it is important to note that the coefficients for Student Gender are nonsignificant in all but one case, suggesting that a student's gender does not differentially affect their ratings of female and male instructors.

Discussion

Previous literature has established evidence of gender bias in SETs (Boring, 2017; Chávez & Mitchell, 2020; Mengel et al., 2019), which raises concerns about the use of these tools as reliable indicators of teaching effectiveness (Hoorens et al., 2021). Our paper sought to further examine *when* and *why* gender bias in SETs occurs by examining SETs for gender-neutral, agentic, and communal qualities at two non-traditional time points in the semester. In accordance with previous literature, we anticipated that bias against female faculty in economics would be present from the onset of the semester due to backlash against women who occupy gender incongruent roles (i.e., a woman being in a position of power and status in a male-dominated field) and would widen over time because of backlash for behaving in gender incongruent ways (i.e., providing feedback).

First, drawing on social role theory (Eagly & Wood, 2016), role congruity theory (Eagly & Karau, 2002; Rudman & Phelan, 2008), and the status incongruity hypothesis (Rudman et al., 2012), we anticipated that women in a high-status, faculty role would experience role-based backlash. This role-based backlash would be expressed as lower evaluations scores on the second day of class for female instructors compared to male instructors. These initial differences in evaluation would be driven by female faculty violating gender role expectations held by students. More specifically, we predicted we would see a significant gender difference across all qualities assessed at Time 1 (H1 a, b, and c), a difference that in the presence of limited student exposure to the professor we would interpret as role-based backlash.

While we found significant instructor gender differences for two of the three gender-neutral items (i.e., *Recommend Instructor*, *Interesting*), and one of two agentic items (i.e., *Challenging*), there were no significant differences for the communal qualities (*Caring*, *Approachable*) at Time 1. However, these differences did not remain after controlling for other potentially explanatory factors. As such, we conclude that our results provide little to no support for role-based backlash occurring at the onset of the semester (Time 1).

We considered two potential explanations for this finding. First, the role of a college faculty member, irrelevant of gender, may not be perceived as a high-status position, which SIH states is critical for backlash to occur (Rudman et al., 2012). While research has indicated economists are perceived as holding a high-status role (Lippa et al., 2014), we do not know if students would perceive faculty as economists, and if this status does not transfer to the faculty role, this would diminish role-based backlash (Rudman et al., 2012). Additionally, because female economics faculty find themselves in a broader industry that is women dominated (i.e., education; U.S. Department of Labor, Bureau of Labor Statistics, 2020), students' perceptions of women in a gender incongruent role might be diluted (U.S. Department of Labor, Bureau of Labor Statistics, 2020).

Second, based on the SIT proposition that women in high-status positions receive more severe backlash when their behavior also violates gender norms and hierarchy (Rudman et al., 2012), we anticipated that male instructor and female instructor ratings would significantly differ after providing exam feedback (H2a, b, c). Additionally, we predicted that the difference between male and female faculty ratings at Time 2 would be significantly larger than the difference between male and female faculty at Time 1 (H3a, b, c). In other words, we predicted that women engaging in gender incongruent behavior that highlights their authority and power in the classroom (i.e., providing feedback), would result in exacerbated student behavior-based backlash evident by even greater differences in ratings on gender-neutral, agentic, and communal qualities than we found at Time 1.

Consistent with Hypothesis 2a, we found significant instructor gender differences at Time 2 for the three gender-neutral items (i.e., *Recommend Instructor*, *Recommend Course*, and *Interesting*), and consistent with Hypothesis 2b for both agentic items (i.e., *Challenging*, *Knowledgeable*), but in contrast to Hypothesis 2c, there were no significant differences for the communal items (i.e., C*aring*, *Approachable*). Importantly, the differences we found remained when controlling for other potentially explanatory factors. As such, we conclude that our results provide support for behavior-based backlash which occurred at a later point in the term after feedback on the first exam was provided (Time 2).

While Hypothesis 2a, 2b, and 2c aligned with our expectations, we found limited support for Hypothesis 3a, 3b, and 3c. While there were larger instructor gender differences across time for all seven qualities, the differences in the instructor gender difference from Time 1 to Time 2 were not significant when controlling for other factors. Despite this, two notable patterns provide support for the premise of Hypothesis 3. First, we found that male

instructors significantly increased from Time 1 to Time 2 on *Recommend Instructor*, *Knowledgeable*, *Challenging*, and *Caring*, while female instructors had no significant change except a significant decrease in *Interesting*. While we expected differences between male and female instructors at Time 2 versus Time 1 to be driven by a significant decrease in evaluations for female faculty, the pattern we find may still be indicative of exacerbated behavior-based backlash. Upon further reflection, it is not surprising that increased exposure over time would lead students to express more favorable attitudes towards all faculty (i.e., the "familiarity breeds liking" effect; Schneider et al., 2012). This familiarity phenomenon did not appear for both male and female faculty. This suggests that either (a) female faculty are not receiving this increase in the same way male faculty are or (b) any increase female faculty may receive is being offset by the backlash for providing negative feedback, which male faculty do not experience. The fact that we see significant increases for male faculty, and virtually no difference for female faculty may still be indicative of the backlash we anticipated to find.

Second, while we did not find significantly larger differences between male and female faculty at Time 2 than we did at Time 1, we did find five of the seven qualities were significantly different between male and female faculty at Time 2, which we did not find at Time 1. The fact that significant differences were found at Time 2 aligns with the idea put forth by SIT that more severe backlash occurs following behavior perceived to be gender incongruent. This aligns with past work demonstrating that even when receiving identical feedback, female managers' feedback is perceived to be significantly less accurate and less appropriate than male managers (Abel, 2019). Moreover, it is consistent with research showing that students perceive women professors to be more incompetent or unfair following the receipt of critical feedback (Sinclair & Kunda, 2000).

Collectively, these results provide insight on when and why gender bias in SETs are occurring and contribute to the growing body of evidence for gender bias in SETs (Boring, 2017; Chávez & Mitchell, 2020; Mengel et al., 2019) and the use of this tool for decision making in higher education (Hoorens et al., 2021). Social role theory and role congruity theory suggest that gender incongruent roles and behaviors will result in the experience of backlash (Eagly & Wood, 2016; Rudman et al., 2012), and our results are generally consistent with a pattern of backlash against female economics instructors.

Limitations and Future Research Directions

As is true with all empirical work, there are limitations that are important to highlight and should be addressed in future research. First, while our participant count was high, one limitation of our sample was the low instructor count. Although comparable studies may have much larger samples of faculty, these studies pool various SET structures, delivery methods, and wordings across course level, course subject, and fields all with their own gender norms and distributions of faculty. Our sample traded high volume for comparable data. Identical teaching evaluations were given to each student for direct comparability across institutions, and students were all surveyed at the exact same two points in time: the second day of class and the day immediately following the first exam's grade reveal. No other study that we are aware of uses this type of comparable data to examine this question. Additionally, we sought to ensure we captured a representative sample coming from a variety of geographic locations (e.g., South, Midwest) and a variety of school types (i.e., universities, liberal arts colleges).

A second, related limitation of the current study is the lack of demographic diversity among our faculty members. While we were intentional in seeking diversity among our faculty, we were unable to find many faculty to commit to the rigorous requirements we had for data contributors. The primary goal of this study was to highlight gender differences; other important demographic factors such as race were beyond the scope of this study. However, it will be important for future work to seek to replicate this approach focusing on the effect of professor race.

A third limitation of the present work is that we have only examined this question in one field: economics. We focused solely on economics to avoid the confounding variable of differences in course level or subject material. Economics is also viewed as a male-dominated field where women occupying these positions would be viewed as more of a gender role violation compared to other more gender balanced fields. Additionally, individual teaching methods and course content is comparable, and most students have to take the course as a requirement. Future work should seek to determine the extent to which backlash is present in other fields – particularly in other male-dominated fields. It would also be beneficial to compare male-dominated disciplines to female-dominated disciplines to determine the extent to which gender role incongruity backlash can explain the differences seen across disciplines. Relatedly, our categorization of traits as agentic or communal was based solely on the existing literature– rather than on the actual impressions of those we surveyed. Future work should seek to assess more directly the impressions that students have regarding the gender typicality of traits for male and female faculty, which will allow us to confirm whether these perceptions are changing over time.

Lastly our paper used a quasi-experimental field data approach rather than a lab-based vignette study. While our

field-based approach does allow for greater generalizability than a lab-based approach, there are limitations of a field-based approach, namely, controlling for confounding variables. We sought to address this by using regression analysis to control for participant effects which allowed us to demonstrate the validity of our conclusions. Future lab-based work to corroborate our findings would be useful.

Practice Implications

Despite some limitations, the present work has advanced both our understanding of when and why gender bias in SETs exists. This work has implications for how colleges and universities may seek to mitigate gender bias in their SETs as well as implications for broader organizational policies on performance evaluations. First, to our knowledge, this is the first study to test the assumed mechanism of backlash directly within a higher education field setting. From an empirical standpoint, this work provides further support and legitimacy to the utilization of this framework by which to understand and examine gender bias within a higher education context. Further, it is possible the results of this study can inform those conducting evaluations of performance in other industries. Within a higher education context, this work provides additional evidence to the growing body of work that suggests bias – as compared to actual differences in teaching effectiveness – accounts for gender differences seen in SETs (Boring, 2017; Chávez & Mitchell, 2020; Mengel et al., 2019). From a practical standpoint, this serves yet as one additional piece of evidence that should call into question the extensive reliance upon SETs in tenure and promotion decisions. Practical strategies would include limiting the role SETs place in these decisions as other studies have also recommended (Hoorens et al., 2021). In the very least, the evidence we find for backlash following feedback from an exam may also suggest it is important to consider the timing of SETs, perhaps holding them prior to final exams, for example.

We also believe the results of our paper offer support for the use of significance testing when comparing evaluation scores between faculty when using SET data for merit-based evaluation of faculty. As researchers, we value and recognize the importance of significant findings. It is important to note that in the real-world application of these data, small, and at times, nonsignificant, differences in means are used to make decisions around selection, pay, promotion, and tenure decisions (Martin, 1984; Nowell et al., 2010). Due to the tight distribution of course evaluation scores among faculty, any differences, though commonly small and often not statistically significant, are used to make consequential decisions (Boysen, 2015). When colleges and universities collect teaching evaluation data, rarely if ever are there controls to ensure representative samples, nor controls for type or size of class or subject, and significance testing is not commonly performed (Becker & Watts, 1999). As such, small, and even nonsignificant differences in course evaluations can adversely affect female faculty in the academic market. While some institutions have taken steps to perform tests of significance, this is a step that our evidence suggests more colleges and universities should take. Additionally, future works should focus on how non-significant differences actually influence decision makers in departments.

While we encourage colleges and universities to use SETs in a limited fashion and with caution, we recognize that a complete abandonment of this measurement tool is unlikely. As such, we hope that the present work serves as further confirmation that bias – likely operating at the unconscious level –drives these differences. Just as managers have long been taught to recognize the role bias plays in their evaluations of female employees, the potential exists for this same form of instruction to be presented to students.

Conclusion

In an effort to better understand the reasons why women remain underrepresented as faculty – particularly as tenured faculty – we sought to test the presence of gender bias at two non-traditional points in time for SETs – a factor widely considered to contribute to gender differences among tenured faculty. We found limited empirical evidence of bias at the beginning of the semester which represents a point in time where students have little basis for rating male and female faculty differently; however, we found clear evidence of bias immediately after the first exam grade is returned. Collectively this work advances our understanding of gender bias in SETs by allowing us to better understand when – and why – students may respond to their female faculty in biased ways to better manage the impact this bias can have on women's ability to thrive.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11199-022-01299-w.

Author Contribution The work provided in this manuscript is wholly collaborative and each author contributed to the overall result.

Funding This study was not funded in any way.

Availability of Data and Material All data is available upon request in summary form. IRB compliance restricts our ability to give out individual level data.

Code Availability All STATA code is available upon request.

Compliance with Ethical Standards

Conflicts of Interest None of the authors have any conflicts of interest to disclose in regards to this research.

References

Abel, M. (2019). Do workers discriminate against female bosses? IZA Discussion Paper No. 12611. https://doi.org/10.2139/ssrn.3457655

Al-Bahrani, A., Buser, W., & Patel, D. (2020). Early causes of financial disquiet and the gender gap in financial literacy: Evidence from college students in the Southeastern United States. *Journal of Family and Economic Issues, 41*(3), 558–571. https://doi.org/10.1111/joca.12205

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*(2), 153–166. https://doi.org/10.1023/a:1008168421283

American Association of University Women. (2020, March 27). *Fast facts: Women working in academia*. AAUW. https://www.aauw.org/resources/article/fast-facts-academia/

Arbuckle, J., & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles, 49*(9/10), 507–516. https://doi.org/10.1023/a:1025832707002

August, L., & Waltman, J. (2004). Culture, climate, and contribution: Career satisfaction among female faculty. *Research in Higher Education, 45*(2), 177–192. https://doi.org/10.1023/b:rihe.0000015694.14358.ed

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*(3), 193–210. https://doi.org/10.1080/03634529909379169

Barsh, J., & Yee, L. (2012). *Unlocking the full potential of women at work: Organization*. McKinsey & Company. https://www.mckinsey.com/business-functions/people-and-organizational-performance/our-insights/unlocking-the-full-potential-of-women-at-work

Becker, W., & Watts, M. (1999). How departments of economics evaluate teaching. *The American Economic Review*, *89*(2), 344–349. https://doi.org/10.1257/aer.89.2.344

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

Boysen, G. A. (2015). Uses and misuses of student evaluations of teaching: The interpretation of differences in teaching evaluation means irrespective of statistical information. *Teaching of Psychology*, *42*(2), 109–118. https://doi.org/10.1177/0098628315569922

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some stem fields more gender balanced than others? *Psychological Bulletin, 143*(1), 1–35. https://doi.org/10.1037/bul0000052

Chevalier, J. (2019, December 13). *The 2019 report of the committee on the status of women in the economics profession*. American Economic Association. https://www.aeaweb.org/content/file?id=11672

Chevalier, J. (2020, December 16). *The 2020 report of the committee on the status of women in the economics profession*. American Economic Association. https://www.aeaweb.org/content/file?id=13749

Chávez, K., & Mitchell, K. M. W. (2020). Exploring bias in student evaluations: Gender, race, and ethnicity. *PS: Political Science & Politics*, *53*(2), 270–274. https://doi.org/10.1017/s1049096519001744

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3), 281–309. https://doi.org/10.3102/00346543051003281

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

Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology, 5*(1), 1–35. https://doi.org/10.1080/14792779543000002

Eagly, A. H., & Wood, W. (2016). *Social role theory of sex differences*. The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies. https://doi.org/10.1002/9781118663219.wbegss183

Equal Rights Advocates. (2003). *Creating gender equity in academia: Equal rights advocates' higher education legal advocacy project roundtable report*. University Women. http://universitywomen.law.stanford.edu/reports/Creating_gender_equity_academia.pdf

Felkey, A. J., & Batz-Barbarich, C. (2021). Can women teach math (and be promoted)? A meta-analysis of gender differences across student evaluations of teaching. *AEA Papers and Proceedings, 111*, 184–189. https://doi.org/10.1257/pandp.20211125

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878. https://doi.org/10.1037/0022-3514.82.6.878

Gasser, C. E., & Shaffer, K. S. (2014). Career development of women in academia: Traversing the leaky pipeline. *The Professional Counselor*, *4*(4), 332–352. https://doi.org/10.15241/ceg.4.4.332

Ginther, D. K., & Kahn, S. (2004). Women in economics: Moving up or falling off the academic career ladder? *Journal of Economic Perspectives, 18*(3), 193–214. https://doi.org/10.1257/0895330042162386

Goulden, M., Mason, M. A., & Frasch, K. (2011). Keeping women in the science pipeline. *The ANNALS of the American Academy of Political and Social Science, 638*(1), 141–162. https://doi.org/10.1177/0002716211416925

Hale, G., & Regev, T. (2014). Gender ratios at top PhD programs in economics. *Economics of Education Review, 41*, 55–70. https://doi.org/10.1016/j.econedurev.2014.03.007

Hentschel, T., Heilman, M. E., & Peus, C. V. (2019). The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Frontiers in Psychology, 10*(11), 1–19. https://doi.org/10.3389/fpsyg.2019.00011

Hoorens, V., Dekkers, G., & Deschrijver, E. (2021). Gender bias in student evaluations of teaching: Students' self-affirmation reduces the bias by lowering evaluations of male professors. *Sex Roles, 84*(1–2), 34–48. https://doi.org/10.1007/s11199-020-01148-8

Jonung, C., & Stahlberg, A. -C. (2009). Does economics have a gender? *Econ Journal Watch*, *6*(1), 60–72. https://lup.lub.lu.se/record/1372522

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: An experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication, 35*(5), 603–625. https://doi.org/10.1177/1075547012472684

Krefting, L. A. (2003). Intertwined discourses of merit and gender: Evidence from academic employment in the USA. *Gender, Work and Organization, 10*(2), 260–278. https://doi.org/10.1111/1468-0432.t01-1-00014

Liaw, S. H., & Goh, K. L. (2003). Evidence and control of biases in student evaluations of teaching. *International Journal of Educational Management, 17*(1), 37–43. https://doi.org/10.1108/09513540310456383

Lippa, R. A., Preston, K., & Penner, J. (2014). Women's representation in 60 occupations from 1972 to 2010: More women in high-status jobs, few women in things-oriented jobs. *PLoS One*, *9*(5), Article e95960. https://doi.org/10.1371/journal.pone.0095960

Macfadyen, L. P., Dawson, S., Prest, S., & Gašević, D. (2016). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education, 41*(6), 821–839. https://doi.org/10.1080/02602938.2015.1044421

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

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253–388. https://doi.org/10.1016/0883-0355(87)90001-2

Martin, E. (1984). Power and authority in the classroom: Sexist stereotypes in teaching evaluations. *Signs: Journal of Women in Culture and Society, 9*(3), 482–492. https://www.jstor.org/stable/3173716

Mason, M. A., & Goulden, M. (2004). Marriage and baby blues: Redefining gender equity in the academy. *The ANNALS of the American Academy of Political and Social Science, 596*(1), 86–103. https://doi.org/10.1177/0002716204268744

McDowell, J. M., Singell, L. D., & Ziliak, J. P. (2001). Gender and promotion in the economics profession. *ILR Review, 54*(2), 224–244. https://doi.org/10.1177/001979390105400202

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal, 35*(1), 37–51. https://doi.org/10.1057/palgrave.eej.9050042

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/jeea/jvx057

Misra, J., Lundquist, J., Holmes, E. D., & Agiomavritis, S. (2010). *Associate professors and gendered barriers to advancement*. The University of Kansas, Department of Philosophy, College of Liberal Arts & Sciences. https://philosophy.ku.edu/sites/philosophy.ku.edu/files/docs/mentoring_documents/Gendered%20Barriers%20to%20Advancement.pdf

Nowell, C., Gale, L. R., & Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment & Evaluation in Higher Education, 35*(4), 463–475. https://doi.org/10.1080/02602930902862875

Perry, M. J. (2019, October 9). *Women earned majority of doctoral degrees in 2018 for 10th straight year and outnumber men in grad school 139 to 100*. AEI. https://www.aei.org/carpe-diem/women-earned-majority-of-doctoral-degrees-in-2018-for-10th-straight-year-and-outnumber-men-in-grad-school-139-to-100/

Ridgeway, C. L., & Correll, S. J. (2004). Unpacking the gender system. *Gender & Society, 18*(4), 510–531. https://doi.org/10.1177/0891243204265269

Riegle-Crumb, C., & Humphries, M. (2012). Exploring bias in math teachers' perceptions of students' ability by gender and race/ethnicity. *Gender & Society, 26*(2), 290–322. https://doi.org/10.1177/0891243211434614

Rousu, M. C., Corrigan, J. R., Harris, D., Hayter, J. K., Houser, S., Lafrancois, B. A., Houser, S., Lafrancois, B., Onafowora, O., Colson, G., & Hoffer, A. (2015). Do monetary incentives matter in classroom experiments? Effects on course performance. *The Journal of Economic Education, 46*(4), 341–349. https://doi.org/10.1080/00220485.2015.1071214

Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology, 74*(3), 629–645. https://doi.org/10.1037/0022-3514.74.3.629

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., & Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology, 48*(1), 165–179. https://doi.org/10.1016/j.jesp.2011.10.008

Rudman, L. A., & Phelan, J. E. (2008). Backlash effects for disconfirming gender stereotypes in organizations. *Research in Organizational Behavior, 28*, 61–79. https://doi.org/10.1016/j.riob.2008.04.003

Schneider, F. W., Gruman, J. A., & Coutts, L. M. (2012). *Applied social psychology: Understanding and addressing social and practical problems* (2nd ed.). SAGE Publications, Inc.

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin, 26*(11), 1329–1342. https://doi.org/10.1177/0146167200263002

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles, 41*(7/8), 509–528. https://doi.org/10.1023/a:1018839203698

U.S. Department of Education, National Center for Education Statistics. (2019). *Table 318.30: Bachelor's, master's, and doctor's degrees conferred by postsecondary institutions, by sex of student and discipline division: 2017–18*. . National Center for Education Statistics (NCES). https://nces.ed.gov/programs/digest/d19/tables/dt19_318.30.asp?current=yes

U.S. Department of Education, National Center for Education Statistics. (2020). *Table 318.10: Degrees conferred by postsecondary institutions, by level of degree and sex of student: Selected years, 1869–70 through 2029–30*. National Center for Education Statistics (NCES). https://nces.ed.gov/programs/digest/d19/tables/dt19_318.10.asp

U.S. Department of Labor, Bureau of Labor Statistics. (2020). *Table 11: Employed persons by Detailed occupation, sex, race, and Hispanic or Latino ethnicity*. U.S. Bureau of Labor Statistics. https://www.bls.gov/cps/cpsaat11.htm

Wang, M.-T., & Degol, J. (2013). Motivational pathways to stem career choices: Using expectancy–value perspective to understand individual and gender differences in stem fields. *Developmental Review, 33*(4), 304–340. https://doi.org/10.1016/j.dr.2013.08.001

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

Williams, M. J., & Tiedens, L. Z. (2016). The subtle suspension of backlash: A meta-analysis of penalties for women's implicit and explicit dominance behavior. *Psychological Bulletin, 142*(2), 165–197. https://doi.org/10.1037/bul0000039

Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the course experience questionnaire. *Studies in Higher Education, 22*(1), 33–53. https://doi.org/10.1080/03075079712331381121

Winkler, J. A. (2000). Faculty reappointment, tenure, and promotion: Barriers for women. *The Professional Geographer, 52*(4), 737–750. https://doi.org/10.1111/0033-0124.00262

Winslow, S. (2010). Gender inequality and time allocations among academic faculty. *Gender & Society, 24*(6), 769–793. https://doi.org/10.1177/0891243210386728

Wolfinger, N. H., Mason, M. A., & Goulden, M. (2008). Problems in the pipeline: Gender, marriage, and fertility in the ivory tower. *The Journal of Higher Education, 79*(4), 388–405. https://doi.org/10.1080/00221546.2008.11772108