Basic and Applied Social Psychology

ISSN: 0197-3533 (Print) 1532-4834 (Online) Journal homepage: https://www.tandfonline.com/loi/hbas20

Unpacking Backlash: Individual and Contextual Moderators of Bias against Female Professors

Alexandra N. Fisher, Danu Anthony Stinson & Anastasija Kalajdzic

To cite this article: Alexandra N. Fisher, Danu Anthony Stinson & Anastasija Kalajdzic (2019) Unpacking Backlash: Individual and Contextual Moderators of Bias against Female Professors, Basic and Applied Social Psychology, 41:5, 305-325, DOI: 10.1080/01973533.2019.1652178

To link to this article: https://doi.org/10.1080/01973533.2019.1652178

BASIC AND APPLIED SOCIAL PSYCHOLOGY
2019, VOL. 41, NO. 5, 305–325
https://doi.org/10.1080/01973533.2019.1652178

Routledge
Taylor & Francis Group

Unpacking Backlash: Individual and Contextual Moderators of Bias against Female Professors

Alexandra N. Fisher, Danu Anthony Stinson, and Anastasija Kalajdzic

University of Victoria

ABSTRACT

This research unpacks backlash against female professors by examining how individual characteristics and social context interact to predict student evaluations on RateMyProfessors.com. As predicted, students evaluated female professors in high-status departments more negatively than female professors in low-status departments, and this backlash effect was attenuated when the female professor was "hot." Moreover, backlash was most pronounced for female professors who had been hired more recently and who were tough graders. A follow-up experiment replicated the main findings concerning status and attractiveness and suggested that perceived gender nonconformity may help to explain backlash against female professors.

Bias against women in academia has been well documented. Female professors receive systematically worse teaching evaluations from students compared to their male counterparts (Feldman, Statham, Richardson, & Cook, 1993; Mengel, Sauermann, & Zölitz, 2018; Miller & Chamberlin, 2000; Mitchell & Martin, 2018). This discrimination occurs despite the fact that female and male professors demonstrate equal skill and garner equal investment and outcomes from their students. Experimental evidence also affirms a gender bias in teaching evaluations. For example, students provide worse evaluations of an online instructor when they are led to believe that the instructor is female compared to male (MacNell, Driscoll, & Hunt, 2015).

Feminist scholar and social psychologist Laura Rudman coined the term *backlash* to describe these kinds of negative or even hostile reactions to women in the workplace (e.g., Rudman, 1998), and the phenomenon is hardly unique to academia. Backlash commonly targets women in high-status male-dominated fields (Eagly & Karau, 2002; Heilman & Okimoto, 2007). For example, women in politics are less well liked by voters than men (Okimoto & Brescoll, 2010). Women in managerial roles are perceived to be less rational than their male counterparts (Heilman, Block, & Martell, 1995). Compared to men, women in the military are perceived to lack the motivation and leadership skills necessary for success (Boldry, Wood, & Kashy, 2001). The common characteristic of the women in each of these examples is their violation of gender-role expectations. Traditional gender roles prescribe that women should display traits such as warmth, submission, and friendliness and that they should occupy lower status positions within society. Gender roles also stipulate what women should *not* do—that is, women should *not* pursue high-status roles or engage in overly "masculine" or dominant behavior. Women's social role is so well established in society that people implicitly associate women with "low status" (Rudman & Kilianski, 2000).

According to the *status-incongruity hypothesis* (Rudman, Moss-Racusin, Phelan, & Nauts, 2012), backlash toward women in high-status fields may stem from the perceived incongruity between the low ascribed status of women and the high perceived status of the field. This perceived incongruity may threaten people's belief that the world operates according to a set of just and fair rules, rules that include gender-role expectancies. Violations of such rules provoke feelings of anxiety and confusion that people are motivated to dispel, typically with efforts to reaffirm and reassert the rules that have been violated (e.g., Jost & Banaji, 1994). Thus, when a woman occupies a high-status social position, people react with hostility because such a woman poses a threat to an important belief system that helps them to make sense of their world and their place in it (e.g., Brescoll,

CONTACT Danu Anthony Stinson dstinson@uvic.ca Department of Psychology, University of Victoria, PO Box 1700 STN CSC, Victoria, BC, V8W 3P5, Canada.

Supplemental material for this article can be accessed at www.tandfonline.com/hbas.

Uhlmann, & Newman, 2013). Negative evaluations and discrimination against women in high-status fields, then, may be the result of people's attempts to reestablish traditional gender-role expectancies and the current status quo.

University students are hardly immune to these gender-role expectancies and biases. Students expect their female professors to be warm, nurturing, and sensitive to their needs but do not expect the same of their male professors (Bennett, 1982; Heilman & Okimoto, 2007). Students are more likely to use words such as *genius* or *brilliant* to describe their male professors than their female professors (Schmidt, 2015). Students are also more likely to use the word *teacher*, which implies lower status, to describe a female professor, whereas they are more likely to use the word *professor*, which denotes higher status, to describe a male professor (MacNell et al., 2015). These very same biases probably inform students' biased evaluations of female professors relative to male professors.

Most research to date concerning gender bias in teaching evaluations tests a simple main-effect of gender on evaluations: Women are rated more negatively than men (e.g., Boring, 2017; MacNell et al., 2015). Yet we suggest that a more nuanced consideration of backlash against women in academia is warranted. We propose that understanding backlash against academic women requires a broader view of women's experiences, a view that considers how personal characteristics and contextual factors may interact to predict the degree to which women are ultimately perceived as conforming to or violating gender-role expectations. Therefore, in the present research, we seek to unpack backlash against female professors by examining how individual difference factors such as attractiveness, years of experience, and course easiness interact with social context factors such as the status and gender composition of the academic department to predict student evaluations.

Unpacking backlash

Confirmatory hypotheses

Academia is, on the whole, a high-status field. Yet status hierarchies still exist within universities. Some fields are more highly valued than others. For instance, professors in fields such as Engineering, Computer Science, and Business tend to receive more funding (Caplan-Bricker, 2013; Usher, 2012) and have higher salaries (Chronicle of Higher Education, 2011) than professors in fields such as English, History, and Philosophy. As a result, within the broader social context of academia there are female faculty who are working in higher or lower status departments, and we propose that departmental status will moderate backlash. Specifically, we hypothesize that the oft-observed gender bias in teaching evaluations will be most pronounced for faculty in higher status departments, but it will be attenuated, or even eliminated, for faculty in lower status departments (i.e., *the backlash hypothesis*). There are hints within the existing literature to support this hypothesis. For instance, although consistent gender-bias against women has been documented in high-status departments such as Business and Economics (e.g., Mengel et al., 2018), examinations of gender-bias in lower status departments, like English and other humanities, sometimes fail to observe a substantial bias against women (e.g., Bennett, 1982; Punyanunt-Carter & Carter, 2015). These inconsistencies may be explained if departmental status is taken into account.

Backlash theory also leads us to propose that some female faculty will be more vulnerable to backlash than others because of their personal characteristics. If backlash occurs because women in high-status fields violate gender-role and status expectations, then a woman may be shielded from backlash if she is perceived to possess characteristics that increase her gender-role conformity and/or ascribed status. We propose that physical attractiveness is one characteristic that may achieve both of these goals. Women whose appearance conforms to idealized standards of beauty are perceived to be more feminine (Heilman & Saruwatari, 1979). Conforming to beauty ideals also increases a woman's ascribed status (Kalick, 1988). Indeed, beautiful people are (accurately) perceived to be more popular and successful than less attractive people (Langlois et al., 2000). This boost in ascribed status that beauty affords may help to eliminate the gap between ascribed and achieved status that often fuels backlash. Therefore, we propose that highly attractive female professors will be shielded from status-based backlash because their attractiveness increases their perceived gender-role conformity and/or because their attractiveness affords them an ascribed status that matches their high-achieved status. Put another way, we predict that less attractive female professors will be most vulnerable to backlash (i.e., *the vulnerability hypothesis*).

Potential confounds and exploratory hypotheses

Individual factors such as years of experience and course easiness or contextual factors such as the gender

composition of the department may further moderate the expression of backlash against women in academia.

Years of experience

Because both inexperience and youth may be confounded with perceptions of status and attractiveness (e.g., Jones et al., 1995; Wilson, 2008), we examine whether less attractive female professors are more vulnerable to backlash even when years of experience is statistically controlled. Moreover, we will explore the possibility that inexperience and/or youth may increase women's vulnerability to backlash: The oft-observed gender bias in student evaluations is more pronounced for younger faculty (Mengel et al., 2018).

Gender composition of the department

When more women work in a given field, people perceive that field to have lower status (Crawley, 2014). So it is possible that the effects of departmental status and departmental gender composition are confounded for ratings of female professors. Therefore, we examine whether departmental status predicts student evaluations of female professors even when departmental gender composition is controlled. This analysis will also indirectly control for the gender composition of the student body, because the gender composition of faculty is correlated with the gender composition of students across academic departments (e.g., $r = .42$; Tolbert & Oberfield, 1991, p. 311). We also explore whether departmental gender composition moderates backlash against women in high-status departments. One possibility is that women working in male-majority departments will be more vulnerable to backlash. Such women may be tokenized and thus will be expected to adhere more strictly to traditional gender roles (Kanter, 1977), which could exacerbate backlash. Alternatively, a higher proportion of women within a high-status department may be a particularly salient threat to the male-dominant status quo (e.g., Jost & Banaji, 1994), which could prompt more backlash.

Course easiness

Higher status fields such as Science or Engineering are often perceived to be more difficult than lower status fields, such as the Humanities (Rosen, 2018). So the effects of course easiness and departmental status may be confounded for ratings of female professors. We examine whether status predicts student evaluations of female professors above and beyond the perceived difficulty of the course. We also explore whether course easiness moderates backlash against high-status female professors. Teaching a difficult subject or being a "tough grader" may violate student expectations that female professors should be kind and sensitive to their needs (e.g., Bennett, 1982; Heilman & Okimoto, 2007), and this violation of gender-role norms could exacerbate backlash against tough female professors in high-status departments.

We test each of these confirmatory and exploratory hypotheses in Study 1 by analyzing student evaluations posted on the website RateMyProfessors.com (RMP) and follow-up with an experimental test of our vulnerability hypothesis in Study 2.

"Chili" power: student evaluations on RateMyProfessors.com

When deciding which classes to take and which to drop, college and university students often turn to RMP—a popular website where students anonymously evaluate the quality of their instructors. RMP allows students to get a sense of their potential instructors and other students' experiences in the course (Davison & Price, 2009; Edwards, Edwards, Qing, & Wahl, 2007). However, RMP ratings can inform more than just students' course selections. RMP ratings also predict students' expectations about how they will fare in the course and their likelihood of attending and recommending the course to a friend. For better or worse, RMP ratings can also color students' expectations of their professors (Kowai-Bell, Guadagno, Little, Preiss, & Hensley, 2011).

Students are more likely to form favorable impressions of professors with positive RMP ratings and unfavorable impressions of professors with negative RMP ratings, which in turn affects how they evaluate their professor at the end of the term (Lewandowski, Higgins, & Nardone, 2012). Indeed, there is a strong positive association between RMP ratings and formal teaching evaluations collected by colleges and universities (e.g., $r = .68$; Coladarci & Kornfield, 2007, p. 4; see also Timmerman, 2008). University search committees and administrators may also consult RMP ratings when evaluating the suitability and the quality of potential job candidates or when making tenure decisions for existing faculty (Montell, 2006). In this way, RMP ratings can have very real, and very meaningful, consequences for the career trajectories and professional outcomes of academics.

RMP offers a unique opportunity to examine personal and contextual moderators of backlash against female professors because, unlike traditional teaching evaluations, RMP used to allow students to rate their professors' physical attractiveness. When evaluating professors on RMP, students were given the option of

responding either "yeah" or "um no" to the quality "Hotness." If the majority of students indicated that their professor was "hot," then a red chili pepper would appear next to the professor's name. Although RMP recently stated on Twitter.com that this chili pepper was meant to "reflect a dynamic/exciting teaching style" (Daprile, 2018), the chili pepper feature was removed from the website on June 28, 2018, precisely because it physically objectified professors (Flaherty, 2018). However, we collected the data for the current research before the chili pepper feature was removed, and so we can use it as an indicator of professors' physical attractiveness. Certainly, students expect professors with chili peppers to be physically attractive (Sohr-Preston, Boswell, McCaleb, & Robertson, 2016), and past studies have treated the chili pepper as an indicator of professors' physical attractiveness (e.g., Felton, Mitchell, & Stinson, 2004; Johnson & Crews, 2013).

The current research

To test our confirmatory and exploratory hypotheses, we conducted two studies—one quasi-experiment using RMP data and one experiment conducted on Amazon's Mechanical Turk (MTurk). For Study 1, we collected attractiveness ratings, overall teaching evaluations, and course easiness ratings from RMP for female professors working at the top 40 national universities in the United States, as ranked by *US News Best Colleges Review*. To examine whether the status of a department predicts teaching evaluations, we collected these data for female professors from the four highest paying departments (i.e., Engineering, Computer Science, Business/Economics, and Law) and four lowest paying departments (i.e., English, Philosophy, History, and Art History) across institutions according to Salary.com (2015). We also collected information about the year each professor was hired at their institution by searching their personal websites and online curriculum vitaes. In addition, we visited the departmental websites for each professor and counted the number of female and male faculty members in each department to obtain a measure of the gender composition within each professor's department. For comparison purposes, we also collected the same information for a sample of male professors from the same departments and universities. Because we collected these data before the chili pepper feature was removed from RMP, our study is uniquely situated to test our hypotheses regarding the interaction of departmental status and professor attractiveness on student evaluations of their professors.

In Study 2, we orthogonally manipulate female professors' attractiveness and departmental status to test our vulnerability hypothesis. MTurk workers view profiles and teaching materials for female professors who are either typically attractive or highly attractive and working in a low-status or high-status academic department, and they evaluate the professors using the same rating scales that are used on RMP. So that we can explore potential mechanisms that might explain our vulnerability hypothesis, participants also evaluate the professors' femininity, warmth, competence, and dominance.

Our research has the potential to make important theoretical and practical contributions to the existing research on gender bias in academia and to backlash theory itself. By examining the effects of departmental status on gender bias, our research will help to resolve inconsistencies in the current literature regarding the location and magnitude of gender bias in academia. Our research will also add critical nuance to backlash theory by examining how personal characteristics, such as a woman's attractiveness, may affect her vulnerability to backlash. Moreover, our research will extend backlash theory by teasing apart the contributions of status, departmental gender composition, and teaching behavior to women's likelihood of experiencing backlash. This analysis is particularly valuable considering that these factors are often confounded in past research, which has focused on backlash against women who were actively pursuing high-status positions in male-dominated fields (e.g., Rudman et al., 2012). Above all, our research will demonstrate how examining gender alone as a predictor of bias may be too imprecise—too blunt—a method to detect the finer nuances of bias against women because it does not take into consideration the underlying diversity and inherent contextuality of women's experiences (Bohan, 1993). Instead, a more nuanced approach is needed. By examining how certain personal and contextual factors moderate backlash against women, our research will reveal the underlying complexity of backlash as well as help to identify the women who are most vulnerable to backlash within academia—a necessary requisite for addressing long-standing gender inequalities in the field.

Study 1

Method

Sampling method

Beginning in early 2015, we collected RMP ratings of all female professors from high- and low-status academic departments at each of the top 40 universities

in the United states as listed by *US News Best Colleges Review* (U.S. College News, 2015). For comparison, we later collected an equal sample of men from the same fields at each of the 40 universities.[1] In an attempt to match the number of men to the number of women sampled in each department at each university, we adopted the following strategy: If we collected RMP ratings for 10 female professors in the Stanford University Philosophy Department, then we collected the RMP ratings for 10 male professors in the Stanford University Philosophy Department. However, the number of ratings of male professors in a department did not always match the number of ratings of female professors in a department. In the event that there were more available RMP ratings of male professors than there were ratings of female professors for a particular department, we used an online random number generator to determine which ratings of male professors to collect. For example, if there were 20 RMP ratings for 20 male professors but only 10 RMP ratings of female professors in the Stanford Philosophy Department, the research assistant would generate a list of 10 random numbers between 1 and 20 and then record the ratings for the 10 male professors whose rank order corresponded with the randomly generated list of numbers (e.g., if the randomly generated list of numbers began with the number 7, the research assistant would count down the list of male professors until they reached the seventh male professor and then would record that male professor's rating. If the second randomly generated number was 15, then the research assistant would count to the fifteenth rating of a male professor and record that rating, etc.). In the event that there were fewer ratings of male professors than there were of female professors in a department, we collected as many ratings of male professors as were available.

Sample

Data for 6,434 professors were originally collected from RMP. Afterward, RMP ratings of each professor were cross-checked with their respective university website to confirm that the faculty member was indeed affiliated with the university and department for which the RMP rating was made. The existence of 340 faculty members could not be verified, resulting in a total sample of 6,094 ratings (3,134 women, 2,960 men). Finally, to reduce bias resulting from overly positive or negative ratings made by only one student, we excluded professor ratings if they were both greater than 2 *SD*s above or below the mean and made by only one person. This resulted in the exclusion of 24 ratings of female professors (0.8%) and 24 ratings of male professors (0.8%), leaving a sample of 3,110 ratings of female professors and 2,936 ratings of male professors. After collecting these data, it became apparent that the sample sizes across status and attractiveness conditions were very unequal. The smallest subsample consisted of high-status "hot" female professors ($n = 321$), whereas the largest subsample consisted of low-status "not hot" female professors ($n = 1{,}374$). Because such inequality can compromise experimental power to detect meaningful differences (Cohen, 1988), we randomly sampled 321 ratings from each of the three other conditions to correct this inequality. Essentially, this strategy is similar to those used in other "natural" experiments whereby a "treatment" group (in our case, the high-status "hot" female professors) is compared to randomized control groups. Randomly sampling from each of the three other conditions also allowed us to reduce bias in our comparison groups (Craig et al., 2012). Our final sample consisted of 2,264 ratings (1,284 ratings of female professors).

Measures

Status

We sampled ratings for professors in the four highest salaried academic fields/departments (i.e., high status) and the four lowest salaried academic fields/departments (i.e., low status) in the United States according to Salary.com (2015). The four highest salaried fields were Law, Engineering, Business/Economics, and Computer Science, with an average median salary of USD $122,264 per year for assistant professors. The four lowest salaried fields were English, History, Philosophy, and Art History, with an average median salary of USD $89,189 per year[2] for assistant professors. During data collection we noticed that many institutions did not have distinct Art History departments, so this field was dropped from further data collection and analysis.

Attractiveness

A red chili pepper next to a professor's name on RMP indicated that he or she had been rated as "hot" by the majority of raters. Professors who had a red chili pepper next to their name were coded as 1 (hot), and those who did not have a chili pepper were coded as 0 (not hot).

Teaching evaluations

RMP allows students to rate professors on three categories using a 5-point scale: *helpfulness*, from 1

Table 1. Descriptive statistics for overall quality ratings as a function of gender, attractiveness, and status in Study 1.

Gender	Attractiveness	Status	*n*	*M*	*SD*	*SE*	*Mdn*	10% TM	Skew	Kurtosis
Women	Not hot	Low	321	3.74	.90	.05	3.90	3.79	−.56	−.13
		High	321	3.17	1.11	.06	3.20	3.20	−.23	−.83
	Hot	Low	321	4.38	.63	.04	4.50	4.47	−1.3	1.65
		High	321	4.15	.76	.04	4.30	4.24	−.85	.02
Men	Not hot	Low	245	3.76	.91	.06	3.90	3.82	−.58	−.31
		High	245	3.33	1.12	.07	3.50	3.39	−.44	−.73
	Hot	Low	245	4.36	.57	.04	4.50	4.43	−.81	−0.1
		High	245	4.39	.56	.04	4.50	4.46	−.91	.34

Note. TM = trimmed mean.

(*useless*) to 5 (*very helpful*); *clarity*, from 1 (*confusing*) to 5 (*crystal clear*); and *easiness*, from 1 (*hard*) to 5 (*easy*). RMP averages helpfulness and clarity ratings to form a rating of the professor's *overall quality*. To reduce bias and provide a more reliable estimate of professor quality (Otto, Sanford, & Ross, 2008), we recorded this average *overall quality* rating for each professor. We also recorded the average *easiness* rating for each professor.

Gender

We initially inferred gender from the professors' names and the students' use of pronouns in written comments on RMP. Research assistants also cross-checked the gender of each professor by searching online for information that could reveal gender identity/presentation (e.g., photographs and/or written profiles on personal web pages or university websites). Feminine presenting professors and/or professors who used *she/her* pronouns were coded as female (Female = 0) and masculine presenting professors and/or professors who used *he/him* pronouns were coded as male (Male = 1). None of the professors we sampled used gender-neutral pronouns or self-identified as nonbinary on their professional web pages. We acknowledge that gender presentation, gendered naming, and pronouns are not synonymous with gender identity, and so some gender misclassification may have occurred in our data.

Year hired

Research assistants searched professors' university web pages, curriculum vitaes, and personal websites and recorded the year that each professor received an undergraduate degree and/or the date that the professor was hired at their institution. We were able to locate undergraduate degree information for 856 female professors and year hired information for 905 female professors. Both measures were highly correlated ($r = .75$).

Departmental gender-composition

Research assistants visited the departmental website for each female professor in our sample and recorded the number of female professors and the total number of professors in the department. We then calculated the percentage of female faculty in each department. The overall percentage of female faculty in the sampled fields were as follows: Computer Science = 17%, Engineering = 19%, Business/Economics = 22%, Philosophy = 27%, Law = 36%, History = 37%, and English = 49%.[3]

Results

Data analysis strategy

We used univariate analysis of variance (ANOVA) to test our confirmatory hypotheses. In accordance with this journal's statistical reporting policy, we report the means and standard deviations for each cell, as well as the effect size (η_p^2 and Cohen's *d*), mean difference, and the standard error of the mean difference for each hypothesis test. We report additional descriptive statistics in Table 1. We used hierarchical linear regression to evaluate our exploratory hypotheses concerning easiness, years of experience, and departmental gender composition. For these examinations, we report the standardized betas, associated standard errors, and unstandardized parameter estimates as estimates of effect size (for inferential statistics, see the supplemental materials on the Open Science Framework: DOI:10.17605/OSF.IO/27ATD).

Confirmatory hypotheses

Backlash against high-status women. Recall that we predicted that students would express backlash against high-status female professors. Thus, we expected that students would rate high-status female professors less favorably than either low-status female professors or high-status male professors. To appraise the evidence for these hypotheses, we first examined overall quality ratings ($M = 3.90$, $SD = .96$) as a function of gender (0 = female professors, 1 = male professors) and status (0 = low status, 1 = high status). There was a small

Table 2. Pearson correlations for the variables used in Study 1 exploratory analyses for female professors.

	Status	Attractiveness	Year hired	Gender composition	Easiness
Year hired	−.05	.11	—	—	—
Gender composition	−.67	−.02	.02	—	—
Easiness	−.05	.12	.10	.04	—
Overall quality rating	−.20	.41	.01	.16	.35

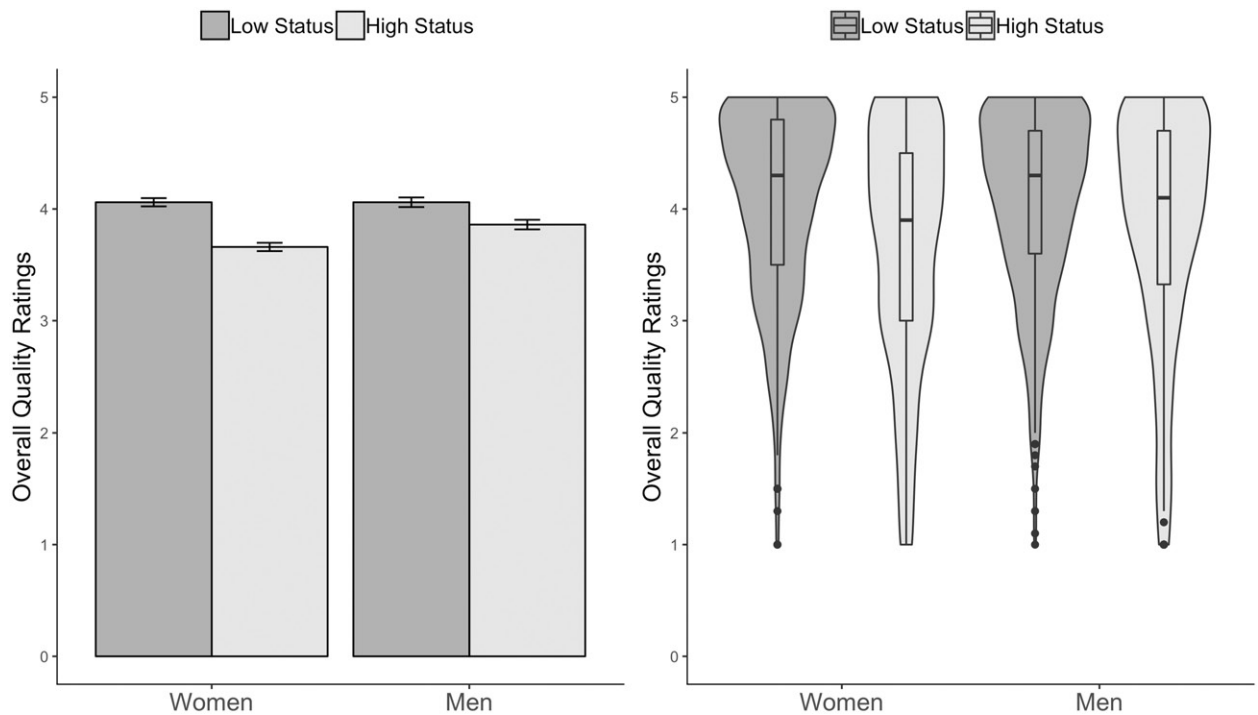

Figure 1. Mean overall quality ratings (left) and their corresponding density distributions and median scores (right) in Study 1 as a function of professor gender and departmental status. *Note:* Error bars represent standard errors.

main effect of gender ($\eta_p^2 = .003$, $d = -.11$): Female professors ($M = 3.85$, $SD = 0.98$) were rated slightly less favorably than male professors ($M = 3.96$, $SD = 0.94$). There was also a main effect of status ($\eta_p^2 = .024$, $d = -0.31$): High-status professors ($M = 3.75$, $SD = 1.06$) were rated less favorably than low-status professors ($M = 4.06$, $SD = 0.83$). There was also evidence of a two-way interaction between gender and status ($\eta_p^2 = .003$, $d = 0.11$; see Figure 1).

Decomposing this interaction revealed backlash against female professors. Students rated high-status female professors less favorably than both low-status female professors ($\eta_p^2 = .024$, $d = -0.31$, *M difference* $= -0.40$, $SE = 0.05$) and high-status male professors ($\eta_p^2 = .006$, $d = -0.16$, *M difference* $= -0.20$, $SE = 0.06$). Gender did not appear to affect the overall quality ratings for low-status professors in this sample ($\eta_p^2 = .00$, $d = 0.00$, *M difference* $= -.002$, $SE = .06$).

Unexpectedly, students also rated high-status male professors less favorably than low-status male professors ($\eta_p^2 = .005$, $d = -0.14$, *M difference* $= -0.20$, $SE = 0.06$), but this status effect was much weaker than the backlash effect observed for female professors. In fact, the status effect for women ($d = -0.31$) was 2.21 times stronger than the status effect for men ($d = -0.14$).

Vulnerability to backlash. We also predicted that less attractive female professors would be more vulnerable to backlash than more attractive female professors. To appraise the evidence for this hypothesis, we examined overall quality ratings for female professors as a function of attractiveness condition (0 = not hot, 1 = hot) and status condition (0 = low status, 1 = high status). In addition to the main effect of status that we have already described, there was a main effect of attractiveness ($\eta_p^2 = .18$, $d = -0.94$), such that the not hot female professors ($M = 3.45$, $SD = 1.05$) were rated less favorably than the hot female professors ($M = 4.27$ $SD = 0.70$). However, there was also

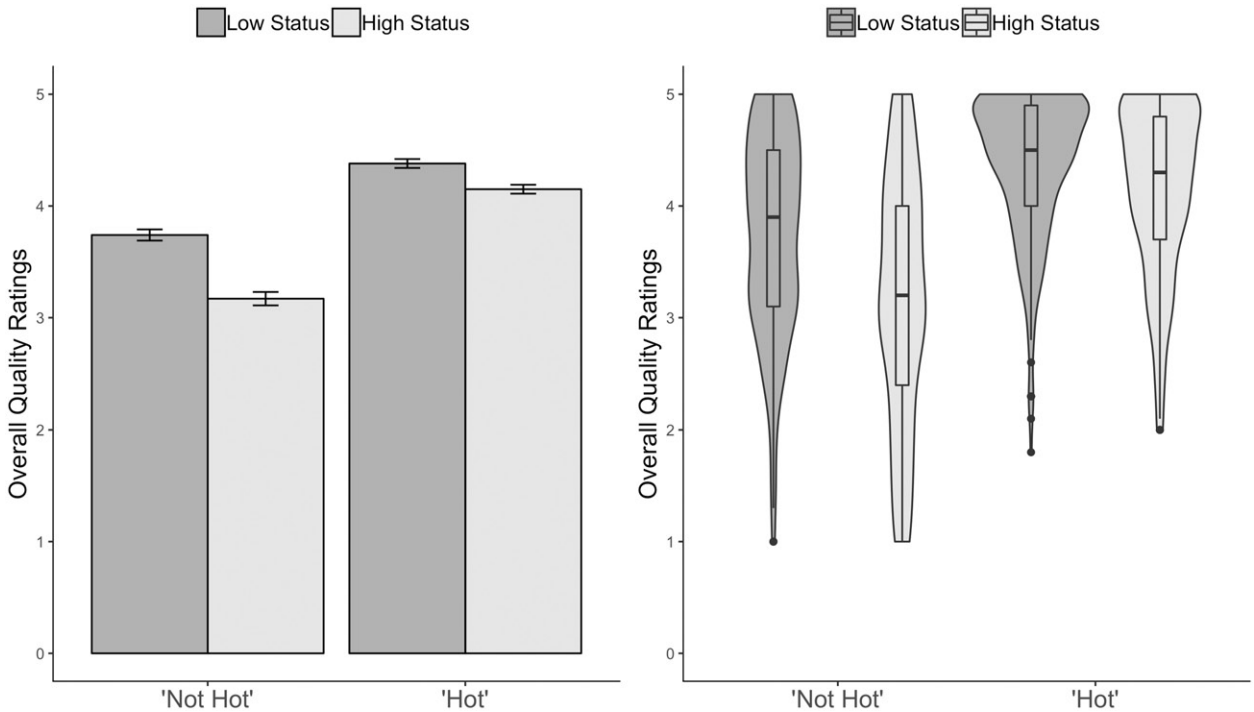

Figure 2. Mean overall quality ratings (left) and their corresponding density distributions and median scores (right) for female professors in Study 1 as a function of professor attractiveness and departmental status. *Note:* Error bars represent standard errors.

evidence for the predicted two-way interaction between status and attractiveness ($\eta_p^2 = .01$, $d = 0.20$; see Figure 2 and Table 1 for descriptive statistics).

Among not hot female professors, students rated women in high-status departments less favorably than women in low-status departments ($\eta_p^2 = .05$, $d = -0.46$, M *difference* $= -0.57$, $SE = .07$). For hot female professors, students also rated women in high-status departments less favorably than women in low-status departments ($\eta_p^2 = .008$, $d = -0.18$, M *difference* $= -0.22$, $SE = 0.07$), but this status effect was much weaker than the observed status effect for not hot female professors. Consistent with our vulnerability hypothesis, the backlash effect for not hot female professors ($d = -0.46$) was 2.56 times stronger than the backlash effect for hot female professors ($d = -0.18$).

Exploratory hypotheses and ruling-out alternative explanations. In each of the following analyses, we first explored whether our confirmatory hypothesis regarding the interaction between departmental status and professor attractiveness remained after controlling for each potential confound (i.e., years of experience, gender composition, course easiness) as well as their interactions with status and attractiveness. We also explored any interactions that emerged between our potential confounds and status. These exploratory analyses allowed us to hone in more precisely on backlash while allowing us to rule out alternative explanations for our observed results (see Table 2 for correlations among variables).

Do years of experience explain our results?

Because age and attractiveness may be confounded (e.g., Jones et al., 1995), and years of experience (a potential proxy for age) and attractiveness were correlated for female professors ($r = .11$), we examined whether status and attractiveness continued to predict female professors' overall quality ratings after controlling for the unique and interactive effects of years of experience. To test this model, we regressed overall quality ratings—Step 1: mean-centered year-hired ($M = 2001.66$, $SD = 9.06$), dummy-coded status (0 = low status, 1 = high status), and dummy-coded attractiveness (0 = not hot, 1 = hot); Step 2: the two-way interactions; and Step 3: the three-way interaction (see the top panel of Table 3 for the results of these analyses). Evidence of the predicted interaction between status and attractiveness remained. There was also evidence of an interaction between year hired and status, such that status-based backlash was stronger for more recently hired women ($\beta = -.42$, $b = -.82$, $SE_b = .11$) than for less recently hired women ($\beta = -.27$, $b = -.52$, $SE_b = .10$). This interaction suggests that less experienced/younger female faculty are more vulnerable to backlash than more experienced/older female faculty.

Table 3. Results of exploratory hierarchical regressions predicting overall quality ratings for female professors in Study 1.

Predictor	*df*	β	*b*	SE_b	ΔR^2	f^2
Year hired						
Step 1	901				.21	.27
Year hired		−.04	−.01	.00		
Attractiveness		.40	.79	.06		
Status		−.29	.45	.06		
Step 2	898				.02	.02
Year Hired × Attractiveness		.08	.01	.01		
Year Hired × Status		−.10	−.02	.01		
Attractiveness × Status		.20	.44	.12		
Step 3	897				.00	.00
Three-way interaction		.09	.02	.01		
Gender composition						
Step 1	1,276				.22	.28
Gender composition		.05	.37	.23		
Attractiveness		.42	.82	.05		
Status		−.17	−.33	.07		
Step 2	1,273				.01	.01
Gender Composition × Attractiveness		.01	.05	.46		
Gender Composition × Status		−.09	−1.04	.46		
Attractiveness × Status		.16	.36	.13		
Step 3	1,272				.00	.00
Three-way interaction		.06	.95	.92		
Easiness						
Step 1	1,280				.30	.43
Easiness		.30	.34	.03		
Attractiveness		.38	.74	.05		
Status		−.19	−.37	.05		
Step 2	1,277				.02	.02
Easiness × Attractiveness		−.14	−.24	.05		
Easiness × Status		.10	.15	.05		
Attractiveness × Status		.11	.26	.09		
Step 3	1,276				.00	.00
Three-way interaction		−.08	−.18	.11		

Does the gender composition in individual departments explain our results?

Because the proportion of female faculty was higher in the low-status departments (43%; $M = .43$, $SD = .11$) compared to the high-status departments (24%; $M = .24$, $SD = .11$) in our sample, and because status and gender composition were strongly and negatively correlated ($r = -.67$), it is possible that the gender composition of departments, rather than status of the departments, accounts for the observed status effects. We tested this possibility, as well as the possibility that departmental gender composition moderates the observed effects of attractiveness and status. We regressed overall quality ratings—Step 1: mean-centered gender-composition ($M = .34$, $SD = .14$), dummy-coded status (0 = low status, 1 = high status), and dummy-coded attractiveness (0 = not hot, 1 = hot); Step 2: the two-way interactions; and Step 3: the three-way interaction (see the middle panel of Table 3 for the results of these analyses). Again, evidence for the predicted interaction between status and attractiveness remained. There was also evidence of an interaction between gender composition and status. In low-status departments, female professors were rated more favorably when there were more women in the department ($\beta = .13$, $b = .89$, $SE_b = .32$), whereas gender composition was not a substantive predictor of ratings of female professors in high-status departments ($\beta = -.02$, $b = -.16$, $SE_b = .33$). This interaction suggests that women working in low-status departments with a higher proportion of female faculty may be less vulnerable backlash.

Does easiness explain our results?

Perhaps our results are not due to backlash at all. Instead, high-status female professors may be tougher graders than low-status female professors, and so it is toughness, not status, that explains overall quality ratings. There was a negligible negative association ($r = -.05$) between status and easiness for female professors in the present study. Still, we regressed overall quality ratings—Step 1: mean-centered easiness ($M = 3.06$, $SD = .85$), dummy-coded status (0 = low status, 1 = high status), and dummy-coded attractiveness (0 = not hot, 1 = hot); Step 2: the two-way interactions; and Step 3: the three-way interaction (see the

bottom panel of Table 3 for the results of these analyses). Evidence for the predicted interaction between status and attractiveness still emerged. There was also evidence of an interaction between easiness and status such that backlash was stronger against tougher female professors ($\beta = -.31$, $b = -.61$, SE_b = .08) than easier female professors ($\beta = -.19$, $b = -.36$, SE_b = .08). This interaction suggests that female professors who teach difficult classes and/or have high academic standards may be more vulnerable to backlash.

Is the attractiveness effect gendered?

Because we observed an unpredicted status effect for men (see Figure 1), we also explored whether status and attractiveness interacted to predict ratings of male professors. Similar to the interaction between attractiveness and status that we described previously for female professors, this follow-up analysis for men also found evidence for a two-way interaction between attractiveness and status ($\eta_p^2 = .02$, $d = 0.29$). Like their female counterparts, students rated not hot male professors in high-status departments less favorably than not hot male professors in low-status departments ($\eta_p^2 = .03$, $d = -0.36$, *M difference* = $-.43$, $SE = .08$). However, this status effect for not hot men was 25% smaller than the observed backlash effect for not hot women ($d = -0.46$). Moreover, status had no discernable effect on ratings of hot male professors ($\eta_p^2 = .00$, $d = 0.00$, *M difference* = .032, $SE = .08$), whereas students still expressed backlash against hot female professors in high-status departments ($d = -0.18$). These results suggest that the observed negative impact of status on RMP ratings for female professors may have a nongendered component that also affects less attractive men, but there is also a gendered backlash that affects women.

Brief discussion

Consistent with our backlash hypothesis, female professors in high-status fields were rated less favorably than male professors in high-status fields and less favorably than female professors in low-status fields. This effect is consistent with the large body of experimental and correlational research documenting a gender bias in teaching evaluations and backlash against high-status women in the workplace (Boring, 2017; MacNell et al., 2015; Rudman et al., 2012). However, our findings concerning the personal and contextual characteristics that moderate academic women's vulnerability to backlash are entirely new to the field. Consistent with our vulnerability hypothesis, backlash was almost twice as severe for not hot compared to hot female professors. This backlash effect was not explained by professors' time at the institution, the easiness of their course, or the gender composition of the department. Of interest, female professors in low-status departments with a higher proportion of female faculty were relatively protected from backlash. However, female professors who were hired more recently or were tough graders were more vulnerable to backlash than those who had been at the institution longer or those who were easier graders.

Because this study used a quasi-experimental design, we cannot conclude that the hypothesized attractiveness effects that we observed are causal. We also cannot rule out potential confounds to explain the observed status effects, including the possibility that the gender composition of the student body is responsible for the backlash we observed against high-status women (i.e., perhaps male students are harsher toward female faculty). Therefore, in our next study we test our primary vulnerability hypothesis experimentally.

Study 2

Our second study uses an experimental method to test our central hypothesis concerning attractiveness and women's vulnerability to backlash in academia. Participants in Study 2 view the ostensible website profile and course materials of a female professor who varies in attractiveness ("hot" vs. "not hot") and status (high vs. low) and then provide evaluations of her teaching ability akin to the ratings made on RMP. We expect to replicate the results of Study 1: Less attractive female professors will be more vulnerable to status-based backlash than more attractive female professors. We also explore whether status and attractiveness interact to predict perceptions of female professors' femininity, dominance, warmth, and competence. Consistent with past research, we suspect that less attractive professors may be more likely to be penalized for being cold, too dominant, or lacking in femininity when they are in high-status departments compared to low-status departments (e.g., Fiske, Xu, Cuddy, & Glick, 1999; Heilman & Okimoto, 2007; Rudman et al., 2012) whereas more attractive professors may be spared from these negative evaluations. This study was preregistered on the Open Science Framework (DOI: 10.17605/OSF.IO/27ATD).

Method

Pilot testing the photographic stimuli

We selected and purchased 10 photographs of women from iStock.com. Each photograph depicted a woman smiling and looking directly into the camera, framed from the waist up. The chosen photographs depicted women who appeared to be between the ages of 45 and 55 and appeared to be from a variety of racial/ethnic backgrounds. Moreover, five photographs were selected because the women conformed very closely to idealized beauty ideals (i.e., "hot" women), and the other five photographs were selected because the women were more typical in appearance (i.e., "not hot" women). We pilot-tested the selected photographs with 161 MTurk workers who rated the women's attractiveness on a 100-point scale from 0 (*not hot*) to 100 (*hot*). Based on the results of this pilot testing, we selected three photographs for the hot condition and three photographs for the not hot condition. In the hot condition, the photographs had a mean attractiveness rating of 66.07 ($SD = 18.05$, $SE = 1.43$), whereas in the not hot condition, the photographs had a mean attractiveness rating of 29.03 ($SD = 20.20$, $SE = 1.61$; $d = 1.93$).

Participants

We recruited 501 participants on MTurk. Based on our preregistered criteria, we excluded participants who did not respond honestly to the survey questions ($n = 39$, defined as a score of 3 or lower in response to the statement, "I tried to answer the questions honestly") and participants who took less than 2 min to complete the survey ($n = 13$). We also removed data from a duplicate IP address ($n = 13$). Our final sample comprised 436 participants (223 men, 2 nonbinary individuals, 211 women; $M_{age} = 35.96$, $SD_{age} = 11.61$; 7% Asian, 4% Black, 7% Hispanic, 78% White). Most participants had at least some experience as a student in a college or university environment ($n = 373$; 85%), and 105 participants (24%) indicated that they were currently a university or college student. Participants were paid USD $1.50 in appreciation for their time.

Procedure

Participants first provided some demographic information and then learned that they would be evaluating a professor from an American university. All participants then viewed a highly realistic webpage profile of a female professor at "Carnegie Mellon University," which included a profile picture and the professor's academic department. Then participants reviewed some of the professor's ostensible teaching materials: a one-page syllabus; two sample lecture slides; and a hand-marked quiz, which included ample written feedback (see the supplemental materials). After viewing the profile and course materials, participants completed the dependent measures. Finally, they were debriefed, awarded their compensation, and thanked for their time.

Independent variables

Physical attractiveness

Each participant was randomly assigned to view either one of the three not hot photographs or one of the three hot photographs. Randomizing target photographs within each condition helps to account for the possibility that some stimuli may elicit higher or lower scores on average than others (Judd, Westfall, & Kenny, 2012).

Status

Participants were randomly assigned to status condition. In the *low-status condition*, the website profile stated that the woman was an associate professor in the *English* department, whereas in the *high-status condition* the profile stated that the woman was an associate professor in the *Engineering* department.

Dependent measures

Professor ratings

Analogous to the rating categories on RMP, participants used a 5-point scale to rate the professor on the following three characteristics: *helpfulness* (1 = *useless*, 5 = *very helpful*), *clarity* (1 = *confusing*, 5 = *crystal clear*), and *easiness* (1 = *hard*, 5 = *easy*). On RMP, helpfulness and clarity are combined automatically into one variable, which we analyzed in Study 1. Accordingly, as stated in our preregistration, we planned to combine these variables. However, these variables could not be reliably combined in the present study ($\alpha = .62$), so we examined clarity and helpfulness separately.

Trait impressions

Participants used a 7-point scale from 1 (*very below average*) to 7 (*very above average*) to rate the professor on several traits, including the following: *femininity* (one item: "feminine"), *warmth* (two items: "warm" and "nurturing"; $\alpha = .81$), *dominance* (one item: "dominant"), and *competence* (two items: "capable" and "competent"; $\alpha = .87$).

Manipulation check

At the end of the survey, participants completed several comprehension items (e.g., "Were you able to

Table 4. Descriptive statistics for primary dependent variables as a function of attractiveness and status in Study 2.

Attractiveness	Status	*n*	*M*	*SD*	*SE*	*Mdn*	10% TM	Skew	Kurtosis
Clarity ratings									
Not hot	Low	114	4.02	.79	.07	4	4.07	−.68	.86
	High	106	3.60	1.04	.10	4	3.67	−.63	−.20
Hot	Low	111	3.93	.86	.08	4	4.01	−.80	.60
	High	104	3.62	1.02	.10	4	3.70	−.83	.28
Helpfulness ratings									
Not hot	Low	113	4.24	.70	.07	4	4.33	−.82	1.04
	High	105	4.12	.77	.07	4	4.21	−.82	.75
Hot	Low	112	4.02	.82	.08	4	4.11	−1.21	2.45
	High	104	4.12	.90	.09	4	4.25	−1.19	1.61

Note. TM = trimmed mean.

Table 5. Descriptive statistics for additional dependent variables as a function of attractiveness and status in Study 2.

Attractiveness	Status	*n*	*M*	*SD*	*SE*	*Mdn*	10% TM	Skew	Kurtosis
Femininity									
Not hot	Low	114	4.96	1.17	.11	5	4.93	.25	−.84
	High	106	4.49	1.30	.13	4	4.53	−.18	−.02
Hot	Low	112	5.37	1.39	.13	6	5.50	−.82	.48
	High	104	5.37	1.39	.14	5	5.45	−.60	.11
Dominance									
Not hot	Low	114	4.74	1.30	.12	5	4.77	−.35	.24
	High	106	4.75	1.32	.13	5	4.79	−.47	.54
Hot	Low	112	4.71	1.47	.14	5	4.84	−.75	.31
	High	104	4.91	1.43	.14	5	5.02	−.84	.64
Warmth									
Not hot	Low	114	4.59	1.21	.11	4.50	4.59	−.05	−.20
	High	106	4.25	1.35	.13	4	4.30	−.27	−.43
Hot	Low	112	4.60	1.27	.12	4.50	4.64	−.28	−.11
	High	104	4.50	1.40	.14	4.75	4.57	−.46	−.24
Competence									
Not hot	Low	114	5.90	.92	.09	6	5.99	−.85	.84
	High	106	5.77	1.05	.10	6	5.89	−1.44	3.43
Hot	Low	112	5.82	.99	.09	6	5.92	−.91	.95
	High	104	5.95	1.14	.11	6	6.14	−1.57	2.96

Note. TM = trimmed mean.

view the professor's profile and course materials [*Yes* or *No*]") and manipulation check items, including one assessing their perceptions of the status of the department (i.e., "How would you rate the status of this department at Carnegie Mellon University"). Participants indicated their responses using a scale from 1 (*very below average*) to 7 (*very above average*).

Results

Data analysis strategy

We adopt the same data analysis strategy that we used for our confirmatory hypotheses in Study 1. We report additional descriptive statistics in Tables 4 and 5, and we present key results in Figure 3.

Manipulation check and preliminary analyses

First we compared participant perceptions of departmental status ($M = 5.64$, $SD = 1.03$) in the high- and low-status conditions. As intended, participants perceived the Engineering department to be appreciably higher in status ($M = 5.77$, $SD = 1.01$) than the English department ($M = 5.54$, $SD = 1.04$, $d = 0.22$, *M difference* = .23, $SE = .10$). Thus, our experimental manipulation was successful, if weaker than we might have desired.

Preliminary analyses revealed that participants' own gender (0 = women, 1 = men, 2 = nonbinary), age, and educational history (0 = never attended college or university, 1 = attended at least some college or university) did not moderate the results we report in a substantive manner, so these variables were omitted from the analyses that follow.

Clarity

Recall that we predicted that less attractive female professors would be more vulnerable to backlash than more attractive female professors. To appraise the evidence for these hypotheses, we first examined clarity ratings ($M = 3.80$, $SD = .94$) as a function of attractiveness condition (0 = hot, 1 = not hot) and status condition (0 = low status, 1 = high status). There was

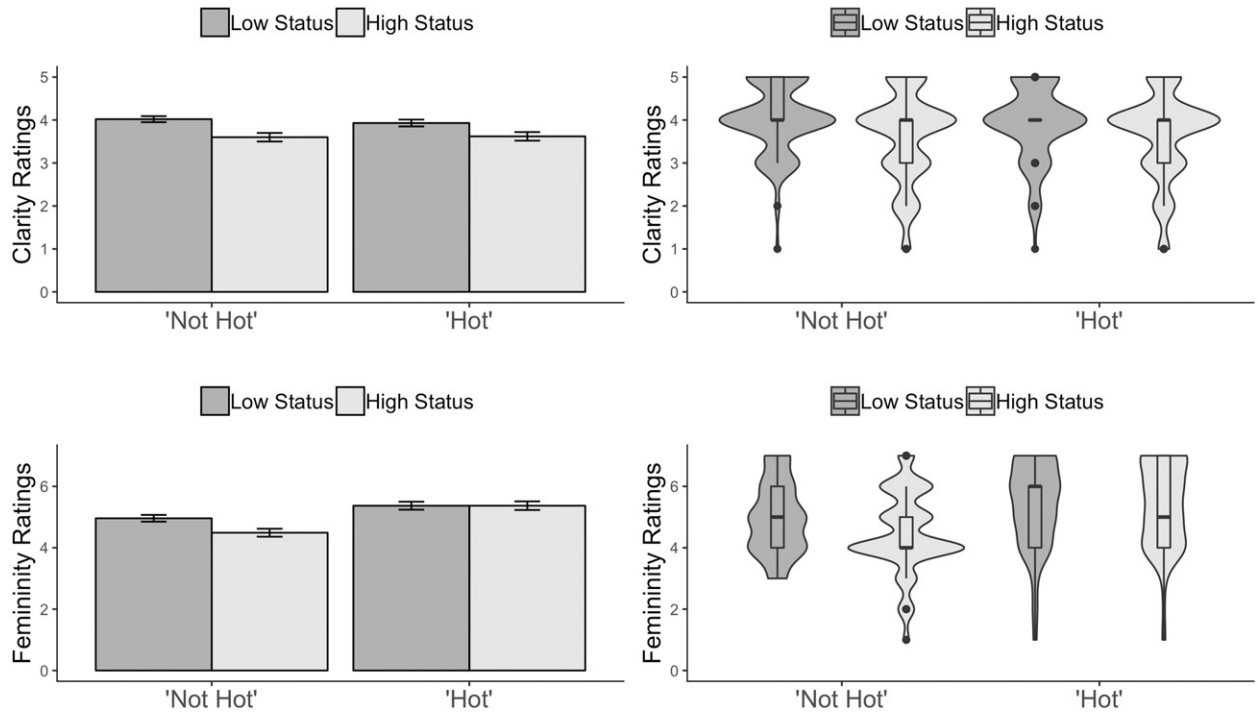

Figure 3. Clarity ratings (top left) and femininity ratings (bottom left) and their corresponding density distributions and median scores (top and bottom right) for female professors in Study 2 as a function of professor attractiveness and departmental status. *Note:* Error bars represent standard errors.

a main effect of status ($\eta_p^2 = .04$, $d = -0.41$): High-status professors ($M = 3.61$, $SD = 1.03$) were rated less favorably than low-status professors ($M = 3.97$, $SD = 0.82$). There was also evidence of a small two-way interaction effect between attractiveness and status ($\eta_p^2 = .001$, $d = 0.06$; see Figure 3).

Among not hot female professors, students rated women in high-status departments less favorably than women in low-status departments ($\eta_p^2 = .03$, $d = -0.35$, *M difference* $= -.41$, $SE = .13$). For hot female professors, students also rated women in high-status departments less favorably than women in low-status departments ($\eta_p^2 = .014$, $d = -0.24$, *M difference* $= -.31$, $SE = .13$), but this status effect was weaker than the observed status effect for not hot women. Consistent with our prediction, the backlash effect for not hot female professors ($d = -0.35$) was 1.45 times stronger than the backlash effect for hot female professors ($d = -0.24$; see Figure 3). In other words, ratings of hot female professors were less affected by the status of their department than were ratings of not hot female professors. As in Study 1, controlling for easiness did not change these results.

Helpfulness

We use our same ANOVA to examine helpfulness ratings ($M = 4.13$, $SD = .79$). There was a small main effect of attractiveness condition ($\eta_p^2 = .005$, $d = 0.14$): Not hot female professors ($M = 4.18$, $SD = .73$) were rated slightly more helpful than hot female professors ($M = 4.07$, $SD = .86$). There was also evidence of a small interaction effect ($\eta_p^2 = .005$, $d = 0.14$).

Among not hot female professors, students tended to rate women in high-status departments slightly less favorably than women in low-status departments ($\eta_p^2 = .003$, $d = -0.11$, *M difference* $= -.11$, $SE = .11$). In contrast, for hot female professors, students tended to rate women in high-status departments slightly more favorably than women in low-status departments ($\eta_p^2 = .002$, $d = .09$, *M difference* $= .11$, $SE = .11$). These small effects contrast with the effects for the combined variable in Study 1 and may be explained by the fact that participants in the current experiment did not actually interact with the professor they rated.

Trait impressions

Femininity

Using the same ANOVA, we examined femininity ratings ($M = 5.05$, $SD = 1.36$) as a function of attractiveness condition (0 = hot, 1 = not hot) and status condition (0 = low status, 1 = high status). There was a main effect of attractiveness ($\eta_p^2 = .06$, $d = 0.51$: Hot female professors were perceived as more feminine

($M = 5.05$, $SD = 1.36$) than the not hot female professors ($M = 4.73$, $SD = 1.26$). There was also a small main effect of status ($\eta_p^2 = .008$, $d = -0.18$): High-status professors ($M = 4.92$, $SD = 1.42$) were perceived to be less feminine than low-status professors ($M = 5.16$, $SD = 1.30$). There was also evidence of a small two-way interaction effect between attractiveness and status ($\eta_p^2 = .008$, $d = 0.18$; see Figure 3).

Participants perceived the not hot female professors to be less feminine when they were in the high-status compared to the low-status department ($\eta_p^2 = .02$, $d = -0.26$, *M difference* $= -.47$, $SE = .18$). In contrast, status did not have an appreciable effect on perceptions of the hot female professors' femininity ($\eta_p^2 = .00$, $d = 0.00$, *M difference* $= -.001$, $SE = .18$). This pattern of effects mirrors the observed effects for clarity.

Dominance

We use our same ANOVA to examine dominance ratings ($M = 4.78$, $SD = 1.38$). There was a small main effect of attractiveness condition ($\eta_p^2 = .001$, $d = 0.06$): Not hot female professors ($M = 4.74$, $SD = 1.30$) were rated slightly less dominant than hot female professors ($M = 4.81$, $SD = 1.45$). There was also a small main effect of status condition ($\eta_p^2 = .001$, $d = 0.06$): High-status female professors ($M = 4.83$, $SD = 1.37$) were seen as more dominant than low-status female professors ($M = 4.73$, $SD = 1.38$). There was also evidence of a small interaction effect ($\eta_p^2 = .001$, $d = 0.06$).

Status did not have an appreciable effect on perceptions of the not hot female professors' dominance ($\eta_p^2 = .00$, $d = 0.00$, *M difference* $= -.008$, $SE = .19$). In contrast, participants tended to perceive the hot female professors as slightly more dominant in the high-status compared to low-status department ($\eta_p^2 = .003$, $d = 0.11$, *M difference* $= .19$, $SE = .19$).

Warmth

We used the same ANOVA to examine warmth ratings ($M = 4.61$, $SD = 1.42$). There was a main effect of attractiveness ($\eta_p^2 = .01$, $d = 0.20$): Hot female professors were perceived as warmer ($M = 4.75$, $SD = 1.45$) than the not hot female professors ($M = 4.47$, $SD = 1.38$). There was also a small main effect of status ($\eta_p^2 = .003$, $d = -0.11$): High-status professors ($M = 4.52$, $SD = 1.47$) were perceived to be less warm than low-status professors ($M = 4.69$, $SD = 1.37$). There was also evidence of a small interaction effect between attractiveness and status ($\eta_p^2 = .002$, $d = 0.09$).

Participants tended to perceive the not hot female professors as slightly less warm when they were in the high-status compared to the low-status department ($\eta_p^2 = .005$, $d = -.14$, *M difference* $= -.29$, $SE = .19$). In contrast, status did not have an appreciable effect on participant perceptions of the hot female professors' warmth ($\eta_p^2 = .00$, $d = 0.00$, *M difference* $= .05$, $SE = .19$).

Competence

Finally, we examined competence ratings ($M = 5.91$, $SD = 1.06$) as a function of status and attractiveness conditions. There was only evidence of a small interaction effect ($\eta_p^2 = .004$, $d = 0.13$). Participants tended to perceive the not hot female professors as slightly less competent in the high-status compared to low-status department ($\eta_p^2 = .002$, $d = -0.09$, *M difference* $= -.13$, $SE = .14$) yet tended to perceive the hot female professors as slightly more competent when they were in the high-status compared to low-status departments ($\eta_p^2 = .002$, $d = 0.09$, *M difference* $= .12$, $SE = .14$).

Brief discussion

The results of this experiment replicated the findings of Study 1, providing experimental support for our proposal that female professors experience status-based backlash that is attenuated by physical attractiveness. It is also notable that the magnitude of the observed effects was similar in both studies. Yet a careful examination of the cell means in each study reveal a subtle difference that is worth exploring in future research. Whereas the stronger backlash effect for not hot female professors in Study 1 appeared to be driven by particularly negative ratings of such women when they were in high-status departments, the stronger backlash effect for not hot female professors in Study 2 appeared to be driven by slightly more favorable evaluations of such women when they were in lower status departments (see Table 4). Thus future research should attempt to determine whether less attractive female professors are vulnerable to backlash because they are rewarded for "staying in their place" in low-status departments, or whether they are penalized for venturing into high-status fields. It is also possible that both of these mechanisms drive backlash against typically attractive women. Of course, whichever way you look at it, less attractive women are more vulnerable to the vicissitudes of status-based bias than more attractive women. Moreover, our exploratory analyses concerning trait ratings of professors suggest that derogation of less attractive women's femininity in high-status fields may indeed be driving

backlash in academia. Consistent with our theorizing regarding the underlying causes of the observed interaction between status and attractiveness, our exploratory analyses concerning trait ratings suggest that perceived gender nonconformity may help to explain backlash against female professors: The female professors who were most vulnerable to backlash (i.e., not hot women working in a high-status department) were rated the lowest in perceived femininity. Once again, these findings help to affirm the gendered nature of backlash.

General discussion

Our findings suggest that status-based backlash is alive and well in RMP ratings. Consistent with our backlash hypothesis, female professors in high-status fields were rated less favorably than female professors in low-status fields. These findings help to resolve inconsistencies in past research by demonstrating that gender bias is stronger, and thus most likely to be detected, in higher status departments such as Business/Economics (Mengel et al., 2018) and weaker, and thus less likely to be detected, in lower status departments such as English (e.g. Bennett, 1982; Punyanunt-Carter & Carter, 2015).

In addition to revealing *where* backlash is most likely to occur, our findings also identify *to whom* backlash is most likely to occur. Consistent with our vulnerability hypothesis, the women who are most at risk of experiencing backlash in high-status departments were those who were considered to be not hot by students (i.e., most female professors). Highly attractive female professors were shielded from the full extent of status-based backlash. Compared to their less attractive counterparts, backlash was attenuated by more than 40% for the highly attractive female professors in Study 1 and by 25% for the highly attractive female professors in Study 2.

Our exploratory analyses in Study 1 revealed some additional personal characteristics that predict greater vulnerability to backlash in academia for women. Specifically, we found that female professors are at greater risk of backlash to the extent that they are less experienced/young, tough graders, and working in high-status departments. In fact, a regression analysis including all of these factors and their interactions predicts that a less experienced not hot woman who teaches a difficult course in a high-status department will receive an average overall RMP rating of 2.57, whereas her more experienced hot counterpart who teaches an easy course in a high-status department will receive an average overall RMP rating of 4.26. This difference represents an effect size of approximately $d = -1.69$, which could translate into lost wages and career opportunities for the most vulnerable group of women. Together, these findings emphasize the utility and importance of employing a more nuanced approach to the study of gender bias and backlash. Had we examined the effect of gender alone on RMP ratings, we would have overlooked the fact that women's experiences of backlash are vastly different depending on their individual characteristics and contexts.

Although the gender composition of the department did not affect female professors in high-status departments, there was a small positive effect of gender composition for women in low-status departments. Female professors in low-status departments were rated better to the extent that there were more women in the department. It is possible that when academic departments hire more women, this creates a more welcoming and inclusive environment that trickles down to influence student evaluations. But it is also possible that benevolent sexism (Glick & Fiske, 1997) leads students to reward female professors who "stay in their place" (i.e., in lower status, lower paid fields) and thus do not threaten the status quo.

Our exploratory findings are also in keeping with Rudman et al. (2012) status-incongruity hypothesis. For instance, newly hired female professors may be more vulnerable to backlash because their youth, inexperience, rank, or any combination of these factors further reduces their ascribed status, resulting in a greater perceived discrepancy between their ascribed status and the high-status of their department. This finding is particularly concerning given that RMP ratings may be used to assess the teaching ability of female academics on the job market or early on in their academic careers (Montell, 2006). In this way, our research suggests one possible reason why academic women continue to be underrepresented in senior positions and continue to be paid less than their male counterparts (Catalyst, 2017)—because backlash against younger/less experienced academic women derails their career progress. Relatedly, although we were limited in the scope of data we could collect about professors, additional factors such as race/ethnicity, weight, religion, and (dis)ability likely interact with other characteristics of women's identities and contexts to predict their experiences of backlash within academia. For instance, past research hints that similar kinds of status-based stereotypes might explain why African Americans and other

minority ethnic and racial groups are especially underrepresented in academia, particularly in fields that ascribe to essentialist notions of status (Storage, Horne, Cimpian, & Leslie, 2016).

Our exploratory findings also suggest that perceived violations of gender roles may increase backlash (e.g., Heilman & Okimoto, 2007). In Study 1, students rated their high-status female professors more negatively to the extent that they violated gender-role expectations by being "tough" graders. Likewise, our exploratory analyses in Study 2 revealed that the women who were most vulnerable to backlash were also deemed to be the least feminine. These findings suggest that both perceived status and behavioral gender-role violations may contribute to backlash against women. Future research should experimentally test the effect of gender-role conformity—specifically femininity—on student evaluations of female professors.

Limitations

One limitation of our Study 1 is that RMP ratings may be biased by self-selection: Students who are especially happy or unhappy may be more likely than other students to submit an evaluation of a professor. That said, self-selection biases also affect actual course evaluations, especially when they are conducted online (Avery, Bryant, Mathios, Kang, & Bell, 2006; Nulty, 2008). Indeed, past research finds that course evaluations tend to be upwardly biased (Goos & Salomons, 2017). Thus, the results of Study 1 might actually provide a fairly accurate reflection of the biases that exist within actual online course evaluations. Nonetheless, we addressed the issue of self-selection in Study 2 by replicating our key findings using an experimental method. Together with Study 2, our field Study 1 allows us to contribute evidence for the ecological validity of backlash research by capturing gender-bias and status-based backlash in a context where it actually occurs (Parigi, Santana, & Cook, 2017), an important contribution to a literature that has relied heavily on hypothetical experimental methods. In this way, our findings enrich and extend existing research that has already demonstrated a high degree of experimental control and internal validity yet has been lacking in practical application.

It is also possible that the majority-male student body in high-status departments was responsible for the backlash effects that we observed in Study 1. After all, young men are more likely to experience threat to their own identity when taught by a high-status woman (Kray, Howland, Russell, & Jackman, 2017; Morton, Postmes, Haslam, & Hornsey, 2009). However, we did not observe participant gender effects in Study 2. This null effect is not definitive. Participants in Study 2 did not know the professors personally, and perhaps more important, they did not receive negative grades from the professor they evaluated. Backlash against female professors is often driven by men (Mengel et al., 2018) and by students who have received low grades (Sinclair & Kunda, 2000). Therefore, more experimental research is needed in an applied setting to determine whether student gender moderates or explains the processes we have documented.

Implications and recommendations

Despite their pervasiveness, student evaluations of teaching (SETs) remain a contentious topic in higher education. Although some research supports student evaluations as a valid measure of teaching quality (Centra, 2003), a large and ever-growing body of research continues to expose the many and varied factors that can bias these evaluations (see Spooren, Brockx, & Mortelmans, 2013, for a review). This literature also calls into question whether SETs truly reflect student learning (Clayson, 2009; Galbraith, Merrill, & Kline, 2012). To complicate matters further, SETs vary widely in format, dimensionality, and purpose (Spooren et al., 2013), which can make it challenging to pinpoint and address the individual factors that lead to bias. Furthermore, some SETs may be more aptly characterized as "customer satisfaction" surveys (Beecham, 2009, p. 135). These types of evaluations, in particular, may be especially conducive to backlash, as they may be more easily swayed by factors that affect students' perceptions of a professors' popularity or status (Clayson, 2009).

Nonetheless, one relatively strong and consistent predictor of SETs is student grades: Students who receive (or expect to receive) good grades are more likely to provide positive course evaluations (Goos & Salomons, 2017). Although this association is often interpreted as evidence for the validity of SETs, such a conclusion may be premature. After all, students who receive poor grades are far more likely to retaliate with negative course evaluations for a female professor than a male professor (Sinclair & Kunda, 2000). The association between grades and SETs also cannot negate the influence of other contextual factors and individual factors that bias student evaluations. For instance, SETs have also been linked to contextual factors such as class size and course workload, as well as

student characteristics such as gender and age (Bedggood & Donovan, 2012; Centra, 2003; Greenwald & Gillmore, 1997; Mengel et al., 2018). At the individual level, professors who are attractive, extraverted, and expressive tend to receive more positive evaluations (Feeley, 2002; Marsh, 1987), as do professors who are seen as approachable or pleasant (Kim, Damewood, & Hodge, 2000) or those who meet a student's expectations of what an "ideal" professor should be (Dunegan & Hrivnak, 2003). Yet, even after accounting for many of these known biasing factors, there remains considerable variance in student ratings left to be explained (e.g., Goos & Salomons, 2017).

Our findings suggest that some of this unexplained variance may be explained by the interaction between contextual factors and the individual characteristics of professors. Specifically, our research suggests that typically attractive, newly hired female professors in high-status departments are especially vulnerable to undeservedly negative evaluations. But we want to be clear about how our research should, and should not, be applied in academic settings. Although our research suggests that the individual characteristics of female professors can explain a meaningful degree of bias in student evaluations, we are *not* proposing that individual female professors are responsible for overcoming backlash. That is, we are *not* suggesting that female professors should change aspects of their appearance, personality, pedagogy/approach, or even their career path to conform to gender-role expectations in an effort to avoid backlash. These kinds of assertions of individual responsibility are a common societal response to structural inequality and do not address the underlying structural problem. For example, the popular "lean in" movement suggests that women can improve their outcomes in high-status workplaces by changing their interpersonal behavior (Sandberg, 2013). Yet social-psychological research demonstrates that such messages have the unintended consequence of blaming women for their own oppression, and thus increase the hostility that such women face in the workplace (Kim, Fitzsimons, & Kay, 2018). Furthermore, women's individual attempts to avoid backlash by moderating their behavior typically fail (Rudman et al., 2012). Thus, the responsibility for combatting backlash against female professors rightfully falls to their academic institutions, who actually hold the power and the means of creating a more equitable work environment. So how can institutions begin to pursue this important goal?

Some researchers recommend that institutions can use SETs for performance evaluation if they statistically correct for student grades and other known biasing factors (McPherson & Jewell, 2007). We disagree. The sheer number of known biasing factors is staggering. Our own research identifies attractiveness, departmental status, years of experience, course easiness, and departmental gender composition as modifiers of backlash. Other research identifies factors like race/ethnicity, cultural background, sexual orientation, and many other factors as sources of bias that influence SETs (e.g. Anderson & Kanner, 2011; Deo, 2015; Fan et al., 2019). It is simply not feasible (or ethical, in the case of some personal/private characteristics) to measure and statistically control for every variable that is known the bias person-perception in general, and SETs in particular. Choosing to control just a few of the most influential factors still leaves other sources of bias that can perpetuate systemic inequity. Instead, we recommend that universities abandon the use of SETs as measures of teaching effectiveness. Biased instruments should not be used to make important hiring, promotion, and tenure decisions at universities. A tempting alternative may be for institutions to create internal committees or hire independent contractors to evaluate teaching quality and effectiveness for performance evaluation. Yet such committees and teams can still harbor bias, even if they have undergone bias-reduction training. Indeed, bias reduction interventions are largely ineffective at reducing bias long term (Lai et al., 2014, 2016), and perhaps consequently, seem to have little impact on changing the proportion of women and people of color in a department (Kalev, Dobbin, & Kelly, 2006). Therefore, we stand by our recommendation that SETs should not be used as measures of teaching effectiveness.

This recommendation does not render SETs useless. Universities and professors may still wish to collect SETs for purposes other than individual performance evaluations. Departmental chairs may wish to examine SETs to identify faculty who may need more support and/or training. Professors may also use SETs to (cautiously) inform their pedagogy and improve their courses. Furthermore, if universities decide to reduce their reliance on SETs, then it is imperative that such institutions provide students with other safe and easy means for reporting concerns about specific professors to the administration (e.g., an ombudsperson). Student must have a say in their education, even if SETs are not the best way to facilitate that process.

Conclusions

Our findings underscore the complexity and social-contextual basis of gender bias in academia. Thus, our research affirms Judith Butler's (1990) warning about the danger of treating women as a "seamless category" (p. 4). By examining the interactions of individual characteristics such as attractiveness, job experience, and course easiness with contextual factors such as the gender composition and status of academic departments, our findings reveal the women who are most vulnerable to backlash—newly hired, typically attractive women teaching difficult courses in high-status departments. Our research also suggests that student evaluations of their professors' teaching ability are biased and should not be used by universities to make employment and tenure decisions about their faculty.

Notes

1. The vulnerability hypothesis was previously labeled *the shield hypothesis* in our preregistration on OSF.
2. This number represents the average salary of history, philosophy, and English professors excluding art history professors according to Salary.com.
3. We also recorded the number of individual ratings that were made for each professor. Accounting for this variable did not alter the results we report.

Acknowledgments

We thank our talented research assistants, Leah Robinson, Amy Rowley, Maria Weaver, Sydney Gelineau-Olay, Nina Perisic, Rachelle Janicki, Maggie Fung, and Anja Hess.

Funding

This research was supported by a Social Sciences and Humanities Research Council of Canada (SSHRC) Master's scholarship to Alexandra N. Fisher and a SSHRC Insight Grant under 435-2014-1579 to Danu Anthony Stinson.

References

Anderson, K. J., & Kanner, M. (2011). Inventing a gay agenda: Students' perceptions of lesbian and gay professors. *Journal of Applied Social Psychology*, *41*, 1538–1564. doi:10.1111/j.1559-1816.2011.00757.x

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *The Journal of Economic Education*, *37*(1), 21–37. doi:10.3200/JECE.37.1.21-37

Bedggood, R. E., & Donovan, J. D. (2012). University performance evaluations: What are we really measuring? *Studies in Higher Education*, *37*(7), 825–842. doi:10.1080/03075079.2010.549221

Beecham, R. (2009). Teaching quality and student satisfaction: Nexus or simulacrum? *London Review of Education*, *7*(2), 135–146. doi:10.1080/14748460902990336

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, *74*(2), 170–179. doi:10.1037/0022-0663.74.2.170

Bohan, J. S. (1993). Regarding gender: Essentialism, constructionism, and feminist psychology. *Psychology of Women Quarterly*, *17*(1), 5–21. doi:10.1111/j.1471-6402.1993.tb00673.x

Boldry, J., Wood, W., & Kashy, D. A. (2001). Gender stereotypes and the evaluation of men and women in military training. *Journal of Social Issues*, *57*(4), 689–705. doi:10.1111/0022-4537.00236

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, *145*, 27–41. doi:10.1016/j.jpubeco.2016.11.006

Brescoll, V. L., Uhlmann, E. L., & Newman, G. E. (2013). The effects of system-justifying motives on endorsement of essentialist explanations for gender differences. *Journal of Personality and Social Psychology*, *105*(6), 891–908. doi:10.1037/a0034701

Butler, J. (1990). *Gender trouble: Feminism and the subversion of identity*. New York, NY: Routledge.

Caplan-Bricker, N. (2013). *New evidence: There is no science-education crisis*. Retrieved from https://newrepublic.com/article/114608/stem-funding-dwarfs-humanities-only-one-crisis

Catalyst. (2017). *Women in academia: Quick take*. Retrieved from https://www.catalyst.org/knowledge/women-academia

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, *44*(5), 495–518. doi:10.1023/A:1025492407752

Chronicle of Higher Education. (2011). *Average faculty salaries by field and rank at 4-year colleges and universities, 2010-11*. Retrieved from https://www.chronicle.com/article/Average-Faculty-Salaries-by/126586

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*(1), 16–30. doi:10.1177/0273475308324086

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coladarci, T., & Kornfield, I. (2007). RateMyProfessors.com versus Formal in-Class Student Evaluations of Teaching. *Practical Assessment, Research & Evaluation, 12*(6), 15.

Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., … Thompson, S. (2012). Using natural experiments to evaluate population health interventions: New MRC guidance. *Journal of Epidemiology and Community Health*, *66*(12), 1182–1186. doi:10.1136/jech-2011-200375

Crawley, D. (2014). Gender and perceptions of occupational prestige: Changes over 20 years. *Sage Open*, 1–11. doi:10.1177/2158244013518923

Daprile, L. (2018). *College students used to rate their professors on 'hotness.' Not any more*. Retrieved Oct 4, 2018 from https://www.thestate.com/news/local/education/article214552980.html

Davison, E., & Price, J. (2009). How do we rate? an evaluation of online student evaluations. *Assessment and Evaluation in Higher Education*, *34*(1), 51–65. doi:10.1080/02602930801895695

Deo, M. E. (2015). A better tenure battle: Fighting bias in teaching evaluations (the more things change: Exploring solutions to persisting discrimination in legal academia. *Columbia Journal of Gender and Law*, *31*(1), 7. doi:10.7916/D8D79PXZ'

Dunegan, K. J., & Hrivnak, M. W. (2003). Characteristics of mindless teaching evaluations and the moderating effects of image compatibility. *Journal of Management Education*, *27*(3), 280–303. doi:10.1177/1052562903027003002

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. doi:10.1037/0033-295X.109.3.573

Edwards, C., Edwards, A., Qing, Q., & Wahl, S. T. (2007). The influence of computer-mediated word-of-mouth communication on student perceptions of instructors and attitudes toward learning course content. *Communication Education*, *56*(3), 255–277.

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), 1–16. doi:10.1371/journal.pone.0209749

Feeley, T. H. F. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education*, *51*, 225–236. doi:10.1080/03634520216519

Feldman, K. A., Statham, A., Richardson, L., & Cook, J. A. (1993). Gender and University Teaching: A negotiated difference. *The Journal of Higher Education*, *64*(1), 117–120. doi:10.2307/2959986

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education*, *29*(1), 91–108. doi:10.1080/0260293032000158180

Fiske, S. T., Xu, J., Cuddy, A. C., & Glick, P. (1999). (Dis)respecting versus (dis)liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of Social Issues*, *55*(3), 473. doi:10.1111/0022-4537.00128

Flaherty, C. (2018). *Bye, bye chili pepper*. Retrieved from https://www.insidehighered.com/news/2018/07/02/rate-my-professors-ditches-its-chili-pepper-hotness-quotient

Galbraith, C., Merrill, G., & Kline, D. (2012). Are student evaluations of teaching effectiveness valid for measuring student outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education*, *53*(3), 353–374. doi:10.1007/s11162-011-9229-0

Glick, P., & Fiske, S. T. (1997). Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly*, *21*(1), 119–135. doi:10.1111/j.1471-6402.1997.tb00104.x

Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: Assessing selection bias in course evaluations. *Research in Higher Education*, *58*(4), 341–364. doi:10.1007/s11162-016-9429-8

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52*(11), 1209–1217. doi:10.1037//0003-066X.52.11.1209

Heilman, M. E., Block, C. J., & Martell, R. F. (1995). Sex stereotypes: Do they influence perceptions of managers? *Journal of Social Behavior and Personality*, *10*(6), 237–252.

Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks? The implied communality deficit. *Journal of Applied Psychology*, *92*(1), 81–92. doi:10.1037/0021-9010.92.1.81

Heilman, M. E., & Saruwatari, L. R. (1979). When beauty is beastly: The effects of appearance and sex on evaluations of job applicants for managerial and nonmanagerial jobs. *Organizational Behavior and Human Performance*, *23*(3), 360–372. doi:10.1016/0030-5073(79)90003-5

Johnson, R. R., & Crews, A. D. (2013). My professor is hot! correlates of RateMyProfessors.com ratings for criminal justice and criminology faculty members. *American Journal of Criminal Justice*, *38*(4), 639–656. doi:10.1007/s12103-012-9186-y

Jones, D., Brace, C. L., Jankowiak, W., Laland, K. N., Musselman, L. E., Langlois, J. H., … Symons, D. (1995). Sexual selection, physical attractiveness, and facial neoteny: Cross-cultural evidence and implications. *Current Anthropology*, *36*(5), 723. doi:10.1086/204427

Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system-justification and the production of false consciousness. *British Journal of Social Psychology*, *33*(1), 1–27. 1994-37261-001

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. doi:10.1037/a0028347

Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American Sociological Review*, *71*(4), 589–617. doi:10.1177/000312240607100404

Kalick, S. M. (1988). Physical attractiveness as a status cue. *Journal of Experimental Social Psychology*, *24*(6), 469–489. doi:10.1016/0022-1031(88)90047-9

Kanter, R. M. (1977). *Men and women of the corporation*. New York, NY: Basic Books.

Kim, C., Damewood, E., & Hodge, N. (2000). Professor attitude: Its effect on teaching evaluations. *Journal of Management Education*, *24*(4), 458–473. doi:10.1177/105256290002400405

Kim, J. Y., Fitzsimons, G. M., & Kay, A. C. (2018). Lean in messages increase attributions of women's responsibility for gender inequality. *Journal of Personality and Social Psychology*, *115*(6), 974–1001. doi:10.1037/pspa0000129

Kowai-Bell, N., Guadagno, R. E., Little, T., Preiss, N., & Hensley, R. (2011). Rate my expectations: How online evaluations of professors impact students' perceived control. *Computers in Human Behavior*, *27*(5), 1862–1867. doi:10.1016/j.chb.2011.04.009

Kray, L. J., Howland, L., Russell, A. G., & Jackman, L. M. (2017). The effects of implicit gender role theories on

gender system justification: Fixed beliefs strengthen masculinity to preserve the status quo. *Journal of Personality and Social Psychology*, *112*(1), 98–115. doi:10.1037/pspp0000124

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., … Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. doi:10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*(3), 390–423.

Lewandowski, G. W., Jr., Higgins, E., & Nardone, N. N. (2012). Just a harmless website?: An experimental examination of RateMyProfessor.com's effect on student evaluations. *Assessment and Evaluation in Higher Education*, *37*(8), 987–1002. doi:10.1080/02602938.2011.594497

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. doi:10.1007/s10755-014-9313-4

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

McPherson, M. A., & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly*, *88*(3), 868–881. doi:10.1111/j.1540-6237.2007.00487.x

Mengel, F., Sauermann, J., & Zölitz, U. (2018). Gender bias in teaching evaluations. *Journal of the European Economic Association*, doi:10.1093/jeea/jvx057

Miller, J., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology*, *28*(4), 283–298. doi:10.2307/1318580

Mitchell, K. M. W., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science and Politics*, *51*(3), 648–652. doi:10.1017/S104909651800001X

Montell, G. (2006). The art of the bogus rating. The Chronicle of Higher Education. Retrieved from http://chronicle.com.ezproxy.library.uvic.ca/article/The-Arts-of-the-Bogus-Rating/46887/

Morton, T. A., Postmes, T., Haslam, S. A., & Hornsey, M. J. (2009). Theorizing gender in the face of social change: Is there anything essential about essentialism? *Journal of Personality and Social Psychology*, *96*(3), 653–664. doi:10.1037/a0012966

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment and Evaluation in Higher Education*, *33*, 301–314. doi:10.1080/02602930701293231

Okimoto, T. G., & Brescoll, V. L. (2010). The price of power: Power seeking and backlash against female politicians. *Personality and Social Psychology Bulletin*, *36*(7), 923–936. doi:10.1177/0146167210371949

Otto, J., Sanford, D. A., & Ross, D. N. (2008). Does ratemyprofessor.com really rate my professor? *Assessment and Evaluation in Higher Education*, *33*(4), 355–368. doi:10.1080/02602930701293405

Parigi, P., Santana, J. J., & Cook, K. S. (2017). Online field experiments: Studying social interactions in context. *Social Psychology Quarterly*, *80*(1), 1–19. doi:10.1177/0190272516680842

Punyanunt-Carter, N., & Carter, S. L. (2015). Students' gender bias in teaching evaluations. *Higher Learning Research Communications*, *5*(3), 28. doi:10.18870/hlrc.v5i3.234

Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: A large-scale study of RateMyProfessors.com data. *Assessment and Evaluation in Higher Education*, *43*(1), 31–44. doi:10.1080/02602938.2016.1276155

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. doi:10.1037/0022-3514.74.3.629

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin*, *26*(11), 1315–1328. doi:10.1177/0146167200263001

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. doi:10.1016/j.jesp.2011.10.008

Salary.com. (2015). *Assistant Professor – English*. Retrieved from https://www.salary.com/

Sandberg, S. (2013). *Lean in: Women, work, and the will to lead* (1st ed.). New York, NY: Alfred A. Knopf.

Schmidt, B. (2015). *Gendered Language in Teaching Reviews*. Retrieved from http://benschmidt.org/profGender/

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. doi:10.1177/0146167200263002

Sohr-Preston, S., Boswell, S., McCaleb, K., & Robertson, D. (2016). Professor gender, age, and "Hotness" in influencing college students' generation and interpretation of professor ratings. *Higher Learning Research Communications*, *6*(3). Retrieved from doi:10.18870/hlrc.v6i3.328

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. doi:10.3102/0034654313496870

Storage, D., Horne, Z., Cimpian, A., & Leslie, S.-J. (2016). The frequency of "brilliant" and "genius" in teaching evaluations predicts the representation of women and African Americans across fields. *PLoS One*, *11*(3), e0150194. Retrieved from http://search.ebscohost.com.ezproxy.library.uvic.ca/login.aspx?direct=true&db=psyh&AN=2016-11626-001&login.asp&site=ehost-live&scope=site

Timmerman, T. (2008). On the validity of RateMyProfessors.com. *Journal of Education for Business*, *84*(1), 55–61. doi:10.3200/JOEB.84.1.55-61

Tolbert, P. S., & Oberfield, A. A. (1991). Sources of organizational demography: Faculty sex ratios in colleges and universities. *Sociology of Education*, *64*(4), 305–315. doi: 10.2307/2112710

U.S. College News. (2015). *National University Ratings*. Retrieved from http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities

Usher, A. (2012). *Average granting council awards per faculty member, by discipline*. Retrieved from http://higheredstrategy.com/research-grants-by-discipline/

Wilson, P. A. (2008). A consideration of tenure and its effect on women faculty members: A proposal for increasing parity between men and women on college campuses. *Forum on Public Policy: A Journal of the Oxford round Table*, 1–33.