GENDER BIAS IN TEACHING EVALUATIONS

by

Kaylyn J. Kim

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Psychological Sciences

West Lafayette, Indiana

August 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Ximena B. Arriaga, Chair
Department of Psychological Sciences
Dr. Janice R. Kelly
Department of Psychological Sciences
Dr. Kipling D. Williams
Department of Psychological Sciences

Approved by:
Dr. David Rollock
Head of the Graduate Program

ACKNOWLEDGMENTS

I would like to thank Dr. Kimberly Kinzig for graciously allowing the use of her online course to be an integral part of this research, Dr. Janice Kelly and Dr. Kipling Williams for their invaluable feedback during the thesis process, and finally, I would like to express my huge gratitude to Dr. Ximena Arriaga for her continuous support and guidance during the entire thesis process as well as my time here at Purdue.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Author: Kim, Kaylyn J. MS
Institution: Purdue University
Degree Received: August 2019
Title: Gender Bias in Teaching Evaluations
Committee Chair: Ximena B. Arriaga

End-of-the-semester teaching evaluations hold consequential weight in professors' career outcomes, which can be problematic if these evaluations are affected by gender bias. This research sought to examine gender bias in evaluations of professors through two experimental studies (via a 15-minute online lecture and a university-sanctioned online course), offering two ecologically valid manipulations of professor gender. Student gender and field of study were examined as moderators of this gender bias, as effects may be more pronounced among male raters compared to female raters, or among raters in majors that underrepresent women compared to raters in other majors. Findings revealed an effect of professor's gender in the opposite direction: On average, students rated female professors more positively than they did male professors. Student gender and field of study did not affect professor ratings, nor did they moderate the effect of professor gender.

INTRODUCTION

Teaching evaluations for university professors are often regarded as a tool of valued insight into the professor's teaching effectiveness and overall course quality. These evaluations hold an important role in determining career outcomes, such as decisions related to hiring, salary, and promotions, including tenure (Abrami, d'Apollonia, & Rosenfield, 2007; Benton & Cashin, 2014). Evaluations also play an essential role in providing feedback to professors on how to improve courses (Chang & McKeachie, 2010), and in the deliberation of committees that select recipients of teaching awards and related recognition opportunities (Benton & Cashin, 2014). Furthermore, a faculty member's career often can be impacted by negative evaluations in other ways, for example, in influencing the decision of promotion and tenure. Given the importance of teaching evaluations, potential bias in these evaluations can have a major impact on who succeeds and who does not beyond actual aptitude and merit.

Gender bias is evident in how students respond in their evaluations, which presents an especially difficult obstacle for female professors (and those of underrepresented groups in general) to overcome in academia. In prior research, much of this gender bias in teaching evaluations has been examined in a naturalistic setting without experimental methods to identify causal variables. For example, previous research has not consistently accounted for factors such as professor age, race, physical attractiveness, gendered teaching style, and gender-dominant subjects, or student-related moderators such as student gender and major. To address these issues, this research involved two experimental studies within a university setting to examine the causal link between professor gender and students' ratings of the professor, while holding professor

features constant (apart from gender), and examining moderators of student gender and major.

Gender Bias in the Workplace

Gender plays an influential role in what others find normative or acceptable in performance, and people often engage in gendered behaviors in order to live up to normative standards and to avoid negative judgments from violating gender assumptions (West & Zimmerman, 1987). In the broader context of Western norms, there is a pervasive devaluation of women in professional settings (Monroe, Ozyurt, Wrigley, & Alexander, 2008) as contrasted with an automatic label of competence that men are granted in professional settings (Johnson, 2006). Workplace settings, which often have deeply-rooted hierarchies of power relations among workers (Ashford, 2003), often impose pervasive gendered expectations and experiences. The arena of higher education is no exception. Female professors face inequality in how students perceive, evaluate, and treat them compared to their male counterparts (Basow, 1995; Centra & Gaubatz, 2000; Feldman, 1993). Women are stereotyped as less logical, less confident, and less competent than men (Acker, 1990), which are all considered vital traits in academia.

When examining gender and the evaluation of leaders, a meta-analysis (Eagly, Makhijani, & Klonsky, 1992) showed a less favorable rating for female leaders and managers in certain circumstances. Specifically, women in leadership positions were devalued relative to their male counterpart when leadership was carried out in stereotypically masculine styles, particularly when this style was autocratic and directive. Additionally, the devaluation of women leaders was greater when leaders were in male-dominated fields and when evaluators were men. Though the context of this research was

in the professional workplace, a similar pattern of teachers as leaders and students as evaluators has been found in academia, such that female professors were more rated more harshly when they failed to meet stereotypic standards of a female.

There has been a myriad of studies on biased gendered judgment in academic evaluative performances. When researchers gave university students identical research articles to evaluate with the one change of a male or female author, students judged the research conducted by the male more highly (Goldberg, 1968; Paludi & Strayer, 1985). In another study in which students evaluated two hypothetical candidates for a faculty position, students tended to judge the male candidate to be more qualified than the female candidate, and sought to hire the male candidate, despite that both candidates held identical credentials (Burns-Glover & Veith, 1995, cf. Moss-Rascusin, 2012). Additionally, because women are expected to be more nurturing and caring, their interpersonal traits are evaluated more critically in evaluation compared to the ratings of their male counterpart. If stereotypic-consistent expectations of warmth and nurturance were not met, women experienced harsher evaluations than men did (Biernat, Fuegen, & Kobrynowicz, 2010; Foschi, 2000).

Teaching Evaluations: Teaching Effectiveness and Interpersonal Traits

In the arena of teaching evaluations, there have been contradictory results regarding main effects on global ratings of professors (Feldman, 1993). When gender-based differences effects on actual teaching evaluations have been found, they have been small or inconsistent (Basow, 1995); sometimes, men receive significantly higher ratings (Basow & Silberg, 1987, Sidanius & Crane, 1989), sometimes, women receive higher ratings (Bachen et al., 1999; Rowden & Carlson, 1996), and sometimes neither (Centra &

Gaubatz, 2000). Because of such inconsistencies, other considerations besides overall or average professor ratings should be taken into account. Specifically, these global ratings may be susceptible to biases operating beyond gender and actual teaching behaviors, such as academic context (e.g., course subject), raters' mood, and professors' physical appearance. In other words, studies that only examine global and overall professor ratings may not adequately reflect the multidimensionality of teaching. The problem is that overall ratings are typically used to assess the essence of a professor's teaching quality, so bias in this measure is especially worth examining.

More nuanced measures of teaching evaluations could detect gender differences in ratings of teaching effectiveness and interpersonal traits (e.g., female professors are consistently rated as being friendlier and having a more positive interpersonal style [Bennet, 1982], as well as facilitating an inviting classroom that fosters feelings of closeness and warmth between students [Crawford & MacLeod, 1990], compared to male professors). This tendency to rate female professors more favorably on interpersonal traits was also reflected in Feldman's (1993) meta-analysis, which concluded that students evaluate female professors more highly on their sensitivity and concerns with class, and on other factors aligned with the interpersonal areas of teaching. Ratings of male professors, however, tended to focus almost exclusively on teaching-effectiveness characteristics, such as confidence, competence, and knowledge of field (Andersen & Miller, 1997). Thus, female professors were uniquely faced with the burden of balancing interpersonal- and competence-related expectations, and those who fail to exhibit an ideal mix of teaching effectiveness and interpersonal traits tended to be rated lower for not

meeting such expectations, a standard to which male professors typically were not held to.

Potential Moderators

The inconsistency in teaching evaluation findings suggests there may be undetected moderating variables that affect evaluations. One moderator may be the gender of the rater. For instance, in actual teaching evaluations, male professors tended to be rated by male and female students fairly consistently, yet female professors tended to be rated lower by male students and higher by female students, a factor that may deceptively show no significant differences in overall ratings (Basow, 1995). Thus, unless student gender is examined in interaction with professor gender, the average ratings of male and female professors may look similar. It has been shown that male students were also less likely than female students to name a female professor as their "best" professor, even when controlling for the total number of female professors they had (Basow, 2000). In contrast, female students often chose female professors as "best" and rated them higher than male professors, especially on qualities related to "fairness" and "providing a comfortable classroom environment" (Basow, Phelan, & Capotosto, 2006).

A second key moderating variable may be the field of study of the student. For instance, it may be the case that professors who do not seem to belong in a given field could receive lower evaluations than ones who do seem to belong. Being a female professor in a female-underrepresented field may result in deductions due to violating expectations of the normative professor in that field. In correlational studies, students' selected majors have been associated with teaching evaluation ratings, such that

engineering majors gave the lowest ratings to non-engineering professors (Basow & Silberg, 1987). A more nuanced understanding of student major and professor gender is necessary, as it may be that students give professors in different areas lower ratings, due to (for example) not being accustomed to that particular area's teaching style or not having a greater appreciation for the subject.

New research using experimental methods can address critical gaps in existing research by experimentally varying the gender of the professor while holding constant both the quality and character of the teaching. Thus, any differences in student ratings can be narrowed to reflecting a bias, while ruling out other causal explanations with as much control as possible. Previous work has demonstrated a complex relationship between ratings of teaching effectiveness and several factors such as student gender, gender-specific discipline, and certain aspects of teaching. Therefore, these moderators have been taken into consideration in the current research.

Perceived Versus Actual Learning

An outcome of interest is whether a professor's gender affects how much students *believe* they have learned from a professor versus how much they *actually* have learned. Student ratings of professors more often reflect how much they believe they have learned rather than their anticipated grades (Baird, 1987). Therefore, perceived learning is an essential component of teaching evaluations. A meta-analysis of learning studies indicated that students' perceived material mastery was weakly correlated with actual material mastery (r = .34), which suggested that students often do not have an accurate representation of how much they actually learned from a course (Sitzmann, Ely, Brown, & Bauer, 2010). Karpicke's work (2011) also indicated that undergraduate students

generally have a poor metacognitive awareness of their perceived learning to actual educational outcomes.

Additionally, there has been a long-contested debate regarding whether perceived learning and actual learning affect teaching evaluations. Students' teaching evaluations may be influenced by the grades they anticipate attaining in a course. When students attained better-than-anticipated grades, they may have perceived that a professor was lenient and thus provided higher teaching evaluations (Greenwald & Gilmore, 1997). Others, however, suggested that better-than-anticipated grades reflect multiple factors, such as background, effort, and amount learned, and may not necessarily be correlated with higher teaching evaluations (Marsh & Roche, 1997). With this discourse, it is apparent that perceived learning and actual learning influence how students respond in teaching evaluations. Therefore, the gap between perceived versus actual learning may be affected by bias. Specifically, students may perceive that they learn less from a female professor when in reality they learn just as much from a female professor as a male professor.

Student gender may also be a factor in perceived learning, as there have been gender differences in causal attributions for imagined academic success and failure. Research on perceived academic success revealed that male students made stronger ability attributions for success than female students did, whereas female students emphasized the importance of studying and paying attention more than male students did, which may shape perceived learning. Male students also attributed failure to a lack of studying and low interest more than female students did, but females were more likely than males to blame failing an exam to a lack of ability (Beyer, 1998). Some of the

gender differences in causal attributions, especially for ability attributions, depended on the gender-type of the subject matter of the examinations, and perhaps the gender of the professor learned from. Therefore, in this research, students' perceived learning was compared to actual learning to observe whether actual differences between male and female students existed.

Methodological Concerns

Previous studies on teaching evaluations examined correlational data, or utilized experimental methods using procedures that lack ecological validity. For example, Basow et al.'s (2013) experimental manipulation of professor gender and race on teaching evaluations offered a computer animation of a professor (manipulated to be a black or white professor and a male or female professor) with a voiceover orating a lecture. This research was high in experimental control (i.e., holding constant factors other than professor gender) but low in realism, as researchers utilized a cartoon to represent a professor, which may not have yielded ecologically valid teaching evaluations. This may have been the reason for inconsistent results in previous literature, such as male students providing the most favorable ratings, as well as African-American professors receiving higher ratings than White professors (Basow, 2013), to correct for bias.

Sinclair and Kunda's (1999) study on gender bias in actual teaching evaluations revealed evidence that students responded differentially to negative feedback from professors, whereby female professors providing negative feedback were evaluated as less competent than male professors providing the same feedback. However, this was not an experimental study, so an experimental approach in a teaching evaluation context would be the next necessary step. Therefore, the current work seeks to improve upon

previous work by using an experimental procedure to manipulate professor gender while maintaining ecological validity by using an actual lecture video and an actual online course.

Current Research

The objective of the current research was to conduct two experimental studies examining the effect of professor gender on student teaching evaluations. While there is a general consensus that gender influences how students perceive and interact with their professors, there is not clear evidence that such a bias affects teaching evaluations. Much of the prior research on gender differences in evaluations has not clearly isolated gender bias as a cause beyond other factors, such as professor's teaching style, speaking style, course subject, physical attractiveness, and age; therefore, these other factors will be held constant. The student's gender and field of study (women-represented major or not) are examined as moderators of interest.

Several *a priori* hypotheses were tested. When an identical lecture is given by either a male or female professor:

> H1: (a) Male professors will be rated higher in teaching effectiveness (knowledge, confidence, competency); (b) a professor gender effect may be moderated by student gender, such that the effect is more pronounced for male raters; and (c) a professor gender effect may be moderated by student major, such that the effect is more pronounced for students in women-underrepresented majors;
>
> H2: Female professors will be rated more highly in terms of interpersonal skills (warmth, kindness, and supportiveness); (b) a

professor gender effect may be moderated by student gender, such that the effect is more pronounced for male raters; and (c) a professor gender effect may be moderated by student major, such that the effect is more pronounced for students in women-underrepresented majors;

H3: Male professors will receive higher overall global professor ratings; (b) a professor gender effect may be moderated by student gender, such that the effect is more pronounced for male raters; and (c) a professor gender effect may be moderated by student major, such that the effect is more pronounced for students in women-underrepresented majors;

H4: Students exposed to a male professor will exhibit higher perceived learning, given that learning may be influenced by teaching effectiveness traits, whereas actual learning scores may not differ by professor gender.

STUDY 1

Study 1 examined gender bias in a 15-minute online lecture, in which participants were randomly assigned to listen to a PowerPoint video led by either a male or female professor. Then, participants took a quiz on the lecture content to measure learning, and then provided evaluations on teaching quality of the professor.

Method

Design

Study 1 was based on a between-subjects design with professor gender (male, female) as a primary independent variable, and two moderating variables: gender of participant (male, female), and field of study (women-represented major, women-underrepresented major). The dependent variables were: teaching effectiveness, interpersonal traits, overall professor rating, perceived learning, and actual learning.

Participants

Participants were 277 undergraduate students (44% male, 56% female, 45% women-underrepresented major, 55% women-represented major) from a large public Midwestern university in the United States. The average age was 19.5 years (*SD* = 1.28 years), and comprised of 60% Caucasian, 25% Asian, 5% African-American, 4% Hispanic, and 6% Other. Power was calculated using G-Power 3.1 for this investigation, and this sample size was determined to have a power of 0.8, which was adequate enough to detect an effect size of 0.2, the effect size used in similar prior research. Participants were recruited through a psychology research participation pool and a psychology course in exchange for research credit or extra credit. All participants were of college age and consented to do the study.

Procedure

The study procedure was conducted through a Qualtrics (online) survey (see the Appendix B for materials). Once informed consent and demographics were obtained, participants were randomly assigned to read about a male or female psychology assistant professor. The page presented a black-and-white headshot of a male or female Caucasian individual in his or her early thirties, along with a 300-word academic biography of the professor. For this study, psychological statistics was chosen as the topic of lecture for its gender-neutral properties, as both male and female undergraduate psychology students report having similar attitudes toward psychological statistics (Walker & Brakke, 2017).

After being presented the professor's photo and biography, participants were directed to a 15-minute PowerPoint video lecture on psychological statistics. The presentation consisted of slides, with a voiceover lecture. The presented lecture was identical in the male and female professor condition, with the one change of a male or female voice to reflect professor gender. Participants were informed that they would take a quiz on the lecture material afterwards and would provide feedback on the professor and the course content. Once all measures were completed, participants were debriefed, thanked, and compensated for their time (via 1 research credit or extra credit).

Measures

Student gender, major, and other demographic characteristics. The survey included items tapping participant gender, ethnicity, age, year in school, and major. Student major was recoded as women-represented or women-underrepresented major based on the faculty gender ratio provided by Purdue Data Digest.

Teaching effectiveness. Teaching effectiveness was measured by ratings of professor traits that are commonly assessed in actual end-of-the-semester evaluations. Participants rated the professor's effectiveness, knowledge in field, confidence, competence, how challenged they felt, and how much they believe they learned from the professor (Cronbach's alpha = .83). All items used a 5-point response scale (1 = *strongly disagree* and 5 = *strongly agree*). The items were averaged into a teaching effectiveness score, and higher numbers indicated more teaching effectiveness traits.

Interpersonal traits. The professor's interpersonal traits were measured using three items (perceived kindness, warmth, and caring; Cronbach's alpha = .91) using a 5-point response scale (1 = *strongly disagree* and 5 = *strongly agree*). The three items were combined into one professor interpersonal traits score. Higher numbers indicated more positive interpersonal traits.

Overall rating. An overall quality rating was obtained by combining two items that are typically used in student teaching evaluations ("Overall, I would rate this instructor as…" and "Overall, I would rate this lecture as…"; 1 = *very poor* and 5 = *excellent*). These items were correlated at r = .83. Higher numbers indicated more positive overall ratings.

Perceived versus actual learning. Perceived learning was measured with a single item: "I felt like I learned a great deal from this instructor" (1 = *strongly disagree* and 5 = *strongly agree*). Higher numbers indicated greater perceived learning. Actual learning was measured using a six-question multiple-choice "quiz" that was created to assess mastery of the lecture content. Participants answered questions by selecting a

correct choice out of four multiple-choice options. These six items were scored by counting the number correct across the six items (score of 0 to 6).

Pilot Study

Pilot data were collected to identify photographs of a male and female to be used in the experimental materials to portray a male and female professor. An independent sample of 194 students evaluated 14 photographs (7 male photos, 7 female photos) in terms of several features: perceived competence, kindness, knowledge in field, intelligence, warmth, supportiveness, physical attractiveness, and age. All ratings used a 7-point rating scale (1 = *not at all* and 7 = *extremely*) for each outcome. The 14 photos were compared to identify one male headshot and one female headshot; the two headshots selected had the most comparable mean levels of perceived competence, kindness, knowledge in field, intelligence, warmth, supportiveness, attractiveness, and age.

Results

To test for the existence of gender bias in teaching evaluations and material mastery, a General Linear Model Univariate analysis was conducted on each dependent variable. For each dependent variable model, an initial analysis included professor gender, student gender, and major, all two-way interactions, and a three-way interaction. None of the three-way interactions were significant and were therefore were dropped. In the second series of analyses, none of the two-way interactions were significant, but the two interactions for hypothesized moderating effects were retained to model their potential influence. Therefore, each model below included three main effects (professor

gender, student gender, major) and the two two-way interactions that were predicted (professor gender x student gender, professor gender x major).

Teaching Effectiveness Traits

For teaching effectiveness ratings (which combined items measuring effectiveness, knowledge, competency, and confidence), a main effect of professor gender (such that the male professor would receive higher scores) was predicted, with male students and women-underrepresented majors rating the male professor higher on teaching effectiveness than female students and women-represented majors would. Results revealed a main effect of professor gender in the opposite way as predicted, such that both male and female students rated the female professor higher on teaching effectiveness than they did the male professor, $F(1, 270) = 4.17$, $p = .04$ (see Table 1). There were no main effects of student gender, $F(1, 270) = .39$, $p = .53$, or major, $F(1, 270) = 1.20$, $p = .27$.

Interpersonal Traits

For professor interpersonal ratings (which combined items measuring warmth, caring, kindness, and supportiveness), a main effect of professor gender (such that the female professor would receive higher scores) was predicted, with male students and women-underrepresented majors rating the female professor higher on interpersonal traits than female students and women-represented majors would. Findings confirmed a main effect of professor gender, $F(1, 272) = 6.15$, $p = .01$, such that female professors received higher ratings on interpersonal traits than did male professors. There were no main effects of student gender, $F(1, 272) = .85$, $p = .36$, or major, $F(1,272) = .01$, $p = .95$.

Overall Rating

For the overall quality rating (which combined items of professor and lecture quality), it was predicted that the male professor would receive a higher score than the female professor would on overall professor rating, with male students and women-underrepresented majors rating the male more favorably than female students or women-represented majors would. Results revealed the opposite pattern, such that the female professor received a higher quality on overall professor rating than the male professor did, $F(1, 272) = 5.88$, $p = .02$. There were no main effects of student gender, $F(1, 272) = .01$, $p = .90$, or major, $F(1, 272) = .31$, $p = .58$.

Perceived vs. Actual Learning

It was predicted that students would have greater perceived learning from the male professor, although actual learning score would be equivalent between professor gender. Perceived and actual learning did not appear to be correlated ($r = -.12$, $p = .31$). Because perceived and actual learning were neither correlated nor conducted on the same scale (perceived learning was measured by a Likert scale from 1-5, and actual learning score a score from 0-6), two separate General Linear Models were conducted on the outcomes. The first General Linear Model was conducted on perceived learning to test whether there would be a student gender, professor gender, or student major effect on perceived learning. Then, a separate General Linear Model was examined on actual learning (which was reflected by the number of correct answers on the quiz) to test whether there would be a student gender, professor gender, or student major effect on actual learning.

Findings revealed that there was not a hypothesized main effect of professor gender effect on perceived learning, as students reported learning similarly from the male and female professor. There was also no main effect of student major on perceived learning. Interestingly, there was a main effect of student gender on perceived learning, such that male students believed they learned more from the lecture and the professor than did female students, $F(1, 270) = 4.78$, $p = .03$.

With respect to actual learning, there were no significant main effects of professor gender, student gender, or student major. Participants performed similarly across conditions on test questions designed to measure content mastery.

Attention

Because expected gender effects were not detected in the analyses, an exploratory quest for other potential moderators was conducted. An attention item ("How closely did you pay attention to the lecture?"; 1 = *not at all* and 5 = *very much*) was examined as a potential moderator, as it could be that students who did not pay attention to the lecture video used more of their gender stereotypes and expectations to complete teaching evaluations. Results indicated that attention to lecture had no main effects of student or professor gender, and it did not influence outcomes of teaching effectiveness, interpersonal traits, or overall professor rating. When comparing the attention item to the actual learning outcome (quiz score), there was a trend in which students who did not pay attention performed worse on the quiz than those who did pay attention, but this effect was not significant, $F(4, 269) = 2.10$, $p = .08$.

Discussion

This study examined the causal patterns that may exist between teaching evaluations and professor gender, taking into account student gender and field of study. Contrary to the hypotheses, Study 1 revealed a pattern favoring the female professor on all teaching evaluation outcomes (teaching effectiveness, interpersonal traits, overall teaching quality). These results suggested that students provide more favorable assessments of female faculty than male faculty, consistent with some (but by no means all or even the majority of) previous work on teaching evaluations (Bachen et al., 1999; Rowden & Carlson, 1996).

A design limitation of Study 1 was that the female lecture used a voiceover by an actual statistics professor, whereas the male lecture used a voiceover by a male graduate student who rehearsed the exact script and nuances conducted by the female voiceover. Although steps were taken to ensure that the male and female video lectures were as identical as possible (such as the male voice actor thoroughly rehearsing the script), it may be that the female voice naturally had an advantage, due to it being narrated by a professor with years of practice delivering a similar lecture that the male actor could not replicate.

Another limitation was that learning did not take place in a realistic classroom setting, in which students and professors interact naturally. In order to examine factors impacting student learning in university environments, an experimental study using a more ecologically valid learning setting was necessary to understand the actual effects of professor gender, student gender, and student major on learning outcomes and professor evaluations. The learning platform of online courses represents an ideal setting, as almost

all professor-related factors could be held constant while still providing a realistic environment.

STUDY 2

Study 2 was conducted in a real-life context: an actual online university course. Although previous studies have indicated that a brief (30-second) nonverbal video clip of a professor, akin to the methodology in Study 1, significantly correlated with the usual end-of-the-semester ratings (Ambady & Rosenthal, 1993), we sought for a deeper confirmation by replicating the study in a controlled and realistic setting of an online course. Study 2 sought to examine gender bias in a university course (Introduction to Behavioral Neuroscience), in which participating students were randomly assigned to a male or female professor who led a portion of the course, when in actuality, the instruction was the same in both groups. Thus, the current research examined the influence of professor gender in both hypothetical and actual contexts, and assessed whether each context would reveal similar patterns of gender bias.

Method

Design

This study design sought to explore key between-subjects independent variables of professor gender (male, female) and the two moderating variables: gender of participant (male, female), and field of study (women represented, women-underrepresented). As in Study 1, the dependent variables included: teaching effectiveness, interpersonal traits, overall professor rating, perceived learning, and actual learning.

Participants

Participants were 36 undergraduate students (34 female, 2 male) who were enrolled in a semester-long online course on Introduction to Behavioral Neuroscience

who completed the teaching evaluation for extra credit. This extra credit was offered to all 55 enrolled students, yielding a 65% participation rate. The average age of students was 20.89 years old (*SD* = 1.30 years), and the ethnic breakdown was 75% Caucasian, 10% Asian, 5% Black, 3% Hispanic, 3% Middle Eastern, and 4% Other.

Procedure

Students who were taking an introductory course on behavioral neuroscience were informed that for the next two weeks, the unit would be taught by a 'visiting professor', who would be interested in feedback on the effectiveness of online teaching styles. The course was carried out completely online, consisting of PowerPoint audio lectures, assigned readings, a discussion post in which the 'visiting professor' provided feedback, and a short quiz. All correspondence between the professor and students was conducted through e-mail. The course was divided into two conditions, and participants were randomly assigned to the one led by a male professor or the one led by a female professor. In actuality, both gender conditions were led by the primary course professor and researcher.

The 'visiting professor' sent out four emails total (see the Appendix B). The first email consisted of an introduction of the professor, including a male or female photo matched to the same biography varying only the male/female pronouns (identical to the profiles used in Study 1), and a link to two lecture videos (with male or female voices) covering the two chapters for that unit. The second email was sent out the first week to correct a mistake on Blackboard in which the discussion post portal was not accessible to students. This mistake was intentional so that students would perceive the visiting professor as only moderately competent (rather than highly competent), in light of

research indicating greater gender bias favoring males when evaluating mediocre, rather than highly competent, candidates (Goldberg, 1968; Paludi & Strayer, 1985). The third email was sent out providing individual feedback to the discussion posts that students submitted a few days prior; the feedback provided to students was identical to all students across conditions. The fourth email was sent out on the second week, which included a reminder to complete assignments, the discussion post, and the upcoming exam. The 'visiting professor' also responded to naturally-occurring student inquiries as they occurred.

After the two-week period was over, the original course professor emailed all students with an 'extra credit opportunity' to answer some questions about the past two weeks' content and to complete a teaching evaluation for the recent visiting professor. Students completed the content quiz, provided teaching evaluations, and then were fully debriefed about the true nature of the study. After being debriefed, students had the opportunity to provide consent for researchers to use their data or have their data completely destroyed. No participant opted to have his or her data deleted.

Measures and Revised Design

The teaching evaluation measures in Study 2 were nearly identical to the one described in Study 1: teaching effectiveness (Cronbach's alpha = .84), interpersonal traits (Cronbach's alpha = .92), overall quality rating (r = .86), perceived learning (single item), and actual learning (0-5 score). There was insufficient variation on student gender (only 2 males completed evaluations) and women-underrepresented majors (3 in this group) to examine these moderating variables. Therefore, the design was modified to include only professor gender as an independent variable.

Results

Because of the nature of the small sample size and the limited representation of student gender and majors, only main effects and general trends of professor gender on outcomes of teaching effectiveness, interpersonal traits, overall teaching quality, perceived learning, and actual learning were examined. Similar to Study 1, a General Linear Model univariate analysis was conducted for professor gender on the dependent variables listed.

Teaching Effectiveness Traits

There were no main effects of professor gender detected in the teaching effectiveness score, $F(1, 34) = 2.23$, $p = .14$.

Interpersonal Traits

There were no main effects of professor gender detected in the interpersonal trait score, $F(1, 34) = .00$, $p = .98$.

Overall Rating

There were no main effects of professor gender detected in overall quality rating, $F(1, 34) = .08$, $p = .78$.

Perceived vs. Actual Learning

There was also no statistically significant within-subjects effects of professor gender on perceived vs. actual learning, $F(1, 32) = .41$, $p = .52$. There were also no main effects of professor gender on perceived learning, $F(1, 34) = 1.88$, $p = .18$; however, there was a marginal difference in actual learning ($F(1, 34) = 3.84$, $p = .06$, such that students scored better with the female professor than they did the male professor (see Table 1).

Attention

There were no main effects of professor gender detected on attention to lecture, $F(1, 34) = .61$, $p = .440$.

Discussion

Study 2 represented a more ecological extension of Study 1. This study examined gender bias in an actual course scenario and explored whether there was stronger evidence of gender bias in this context compared to Study 1. This investigation's strengths lied in its realism, as it used the online learning environment to present a unique opportunity to experimentally manipulate professor gender more directly. Although this study did not replicate the findings found in Study 1, it did offer an interesting slight trend of students reporting to have learned more with the female professor, and performing better on the quiz with the female professor. More generally, it may be that professor gender, student gender, and/or student major may indeed powerfully influence professor ratings, but a higher sample size may be needed to detect these effects.

GENERAL DISCUSSION

Both studies explored whether gender bias affected student ratings of teaching, and whether the continued use of student ratings of teaching as a primary means of assessing quality of professor's teaching may disadvantage women in academia. Study 1 demonstrated evidence of gender bias, but in the opposite pattern anticipated, such that the female professor received more favorable ratings compared to the male professor, despite the two professors presenting identical lectures and controlling for potentially appearance-related variables (e.g. age, physical attractiveness, ethnicity). It is interesting to note that student gender and student major did not play a significant role in how students responded in teaching evaluations. Previous literature indicated moderating effects of student gender, and it was expected that some effect of student gender would influence how students respond to male or female professors, but there was a fairly uniform pattern in how male and female students responded to male and female professors. Similarly, with field of study, students in female-represented majors responded similarly to students in women-underrepresented majors.

In Study 1, the female professor received higher ratings on nearly all outcomes of teaching evaluation, indicating an unexpected pattern of gender bias. It also revealed an interesting pattern of male students having reported learning more than female students did (regardless of what gender their professor was), even if that perceived learning was not reflected in their actual quiz scores. This is inconsistent with previous literature on gender differences in perceived learning in the online environment, which found that female students reported learning more than male students did at the end of the semester (Rovai, 2005). It may be that when faced with a brief 15-minute lecture (instead of a

semester-long course), female students initially felt less confident in their learning ability than males did, although this did not affect performance. However, there is research showing that undergraduate female students were more modest in achievement situations and predicted lower GPAs than male students did, although there was no gender difference in actual GPA (Heatherington et al., 1993). Therefore, further research is needed to examine during which part of the semester timeline or under which academic context do female students regain confidence in their learning ability.

It may also be that using an actual statistics professor as the female voice influenced the quality of the lecture in a way that could not be replicated by the male voice actor. Although the original intention was to use an identical lecture recording and modulate the tone to mimic a male or female voice, it quickly became apparent that the voices sounded unusual, possibly because male and female voices may differ on several facets in addition to the pitch. Therefore, per suggestion of audio editors and voice actors, two separate voiceovers were recorded to reflect a male and female voice.

It was difficult to interpret findings in Study 2 due to limited sample size and sample representation. However, despite these shortcomings, results revealed a slight trend of students reporting more learning from the female professor, and actually performing better on a quiz when having learned from the female professor. This disconfirms the hypothesis on perceived learning, which proposed that perceived learning may be staked on stereotypically-male traits such as competency, confidence, and knowledge in field. This pattern provides a compelling reason to replicate this study in a larger class with a more diverse gender and major representation to gather a better

understanding about the student gender differences in the relationship between perceived and actual learning.

One possible explanation as to why the female professor yielded higher ratings than the male professor may be due to the headshots used to represent each. A pilot study was initially conducted based on fourteen headshots to find identify male and female headshots that are rated with comparable levels of perceived competence, knowledge, kindness, intelligence, supportiveness, warmth, physical attractiveness, and age. However, the analysis of the pilot data was limited to comparing mean levels of each outcome variable for the male and female photos. A subsequent re-analysis of the pilot data using a repeated-measures ANOVA indicated that the female headshot was rated as more physically attractive and older than the male headshot, despite exhibiting similar mean levels on these variables (see the Appendix B). Attractive professors may be rated more favorably in teaching quality than less attractive professors (Bonds-Raacke & Raacke, 2006). Therefore, the female professor may have had an attractiveness advantage that resulted in receiving higher ratings than the male professor. The perception of the female professor as slightly older also may have influence ratings of teaching quality, although the effects of instructor age on teaching evaluations are mixed; younger professors often receiving higher teaching ratings in experimental and correlational studies (Wilson, Beyer, & Montiero, 2014), but older professors are rated to be more competent and knowledgeable (Sohr-Preston, 2016). Future research should make use of male and female photos that are comparable in key ways, including physical attractiveness and age.

Voice may be an untapped factor in how students form opinions about professors (especially in the online learning context) and thus respond in evaluations. Male and female voices are fundamentally different, not only in pitch, but in breathiness, vowel formation, vowel emphasis, and frequency (Price, 1989, Coleman, 1971). For these reasons, it may be that students responded to voice more than to gender, especially in a context in which voice is the most salient feature about an individual. Future research should include identical lecture videos from several different male and female voices to ensure the effects found were not isolated to the voices used in Studies 1 and 2. Moreover, using multiple female and male voices would allow for a test of effects that remain robust across diverse types and ranges of voices.

This study revealed that how gender schemas shape evaluations is not uniform or simple (e.g. males receiving higher ratings of competency and females receiving higher levels of warmth). Bachen et al.'s (1999) naturalistic research on teaching evaluations did find a pattern of female professors receiving higher ratings on global teaching ratings compared to male professors, but this pattern was moderated by student gender, such that female students rated female professors more highly than male professors, whereas male students rated male and female professors similarly. Study 1 in this line of research did find overall higher ratings for female professors, but the student gender moderation was not replicated. It may be that in an experimental setting, male and female students are willing to evaluate professors more objectively than in real life, in which student gender may have a heavier influence on evaluations. Additionally, an online platform may capture differences in teaching evaluations that differ from those in a traditional classroom, as gender bias may not be so prevalent in online courses. It may be that

gender bias is more prevalent in live lectures, in which factors such as hair, clothing, or the mere physical presence of a professor could influence the activation of bias in ways that online lectures do not.

Further research is needed to understand whether and how ratings of teaching involve bias in inconsistent directions, and the legitimacy and merit of such ratings in driving important outcomes (e.g., professor hires, promotions). This work adds to the growing call for re-evaluation and modification of the current system of evaluating the quality of instruction in higher education. For example, students may respond to nuanced and minute differences in vocal mannerisms despite otherwise highly similar voice recordings over and above gender, perhaps providing a partial explanation for the results of Study 1. Additionally, Study 2 should be carried out in online courses with a larger enrollment to examine more accurately associations between student gender and professor gender, as well as student major and professor gender, in teaching evaluations. This study provides important clues for future work that will enhance the design of studies examining gender bias, such as important considerations (e.g., specific preferences for vocal mannerisms) that may drive underlying mechanisms of this bias.

LIST OF REFERENCES

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385-456). Dordrecht, The Netherlands: Springer.

Acker, J. (1990). Hierarchies, jobs, bodies: A theory of gendered organizations. *Gender & Society, 4*(2), 139-158.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64*(3), 431-441.

Andersen, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *PS: Political Science & Politics, 30*(2), 216-219.

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*(3), 193-210.

Baird, J. S. (1987). Perceived learning in relation to student evaluation of university instruction. *Journal of Educational Psychology, 79*(1), 90-91.

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*(4), 656-665.

Basow, S. A. (2000). Best and worst professors: Gender patterns in students' choices. *Sex Roles, 43*(5-6), 407-417.

Basow, S., Codos, S., & Martin, J. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal, 47*(2), 352-363.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30*(1), 25-35.

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*(3), 308-314.

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In M. B. Paulson (Ed.), *Higher education: Handbook of theory and research* (Vol. 29, pp. 279-326). Dorecht, The Netherlands: Springer.

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*, 170-179.

Beyer, S. (1998). Gender differences in causal attributions by college students of performance on course examinations. *Current Psychology, 17*(4), 346-358.

Biernat, M., Fuegen, K., & Kobrynowicz, D. (2010). Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. *Personality and Social Psychology Bulletin, 36*(7), 855-868.

Bonds-Raacke, J. M. (2006). Students' attitudes toward the introduction of a course website. *Journal of Instructional Psychology, 33*(4), 251-256.

Burns-Glover, A. L., & Veith, D. J. (1995). Revisiting gender and teaching evaluations: Sex still makes a difference. *Journal of Social Behavior and Personality, 10*(4), 69-80.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*(1), 17-33.

Chang, T. S., McKeachie, W., & Lin, Y. G. (2010). Faculty perceptions of teaching support and teaching efficacy in Taiwan. *Higher Education, 59*(2), 207-220.

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech and Hearing Research, 14*(3), 565-577.

Crawford, M., & MacLeod, M. (1990). Gender in the college classroom: An assessment of the "chilly climate" for women. *Sex Roles, 23*(3-4), 101-122.

Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. (1992). Gender and the evaluation of leaders: A meta-analysis. *Psychological Bulletin, 111*(1), 3-22.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*(2), 151-211.

Foschi, M. (2000). Double standards for competence: Theory and research. *Annual Review of Sociology, 26*(1), 21-42.

Goldberg, P. (1968). Are women prejudiced against women? *Trans-action, 5*(5), 28-30.

Greenwald, A. G., & Gilmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217.

Johnson, P. (Ed.). (2006). Corporate strategy: Purpose. In *Astute competition (Technology, innovation, entrepreneurship and competitive strategy, Vol. 11)* (pp. 141-155). Bingley, United Kingdom: Emerald Group Publishing Limited.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6018), 772-775.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197.

Monroe, K., Ozyurt, S., Wrigley, T., & Alexander, A. (2008). Gender equality in academia: Bad news from the trenches, and some possible solutions. *Perspectives on Politics, 6*(2), 215-233.

Moss-Rascusin, C. A., Dovidio, J. F., Brescoli, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, USA, 109*, 16474-16479.

Paludi, M. A., & Strayer, L. A. (1985). What's in an author's name? Differential evaluations of performance as a function of author's name. *Sex Roles, 12*(3-4), 353-361.

Price, P. J. (1989). Male and female voice source characteristics: Inverse filtering results. *Speech Communication, 8*(3), 261-277.

Rovai, A. P., & Wighting, M. J. (2005). Feelings of alienation and community among higher education students in a virtual classroom. *The Internet and Higher Education, 8*(2), 97-110.

Rowden, G. V., & Carlson, R. E. (1996). Gender issues and students' perceptions of instructors' immediacy and evaluation of teaching and course. *Psychological Reports, 78*(3), 835-839.

Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology, 19*(2), 174-197.

Sinclair, L., & Kunda, Z. (1999). Reactions to a black professional: motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology, 77*(5), 885-904.

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education, 9*(2), 169-191.

Sohr-Preston, S. L., Boswell, S. 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). DOI: https://doi.org/10.18870/hlrc.v6i3.328

Walker, E. R., & Brakke, K. E. (2017). Undergraduate psychology students' efficacy and attitudes across introductory and advanced statistics courses. *Scholarship of Teaching and Learning in Psychology, 3*(2), 132-140.

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & Society*, 1(2), 125-151.

APPENDIX A

Table 1

Pearson Correlations Among Teaching Evaluation Outcomes

	Teaching Effectiveness	Interpersonal Traits	Overall Rating
1. Teaching Effectiveness	—	—	—
2. Interpersonal Traits	.597**	—	—
3. Overall Rating	.714**	.623**	—

** $p < .01$.

Table 2

Means and Standard Deviations of Teaching Evaluation Outcomes by Student Major, Student Gender, and Professor Gender

	Women-Underrepresented Major				Women-Represented Major			
	Female Rater		Male Rater		Female Rater		Male Rater	
	Female Professor	Male Professor	Female Professor	Male Professor	Female Professor	Male Professor	Female Professor	Male Professor
Teaching Effectiveness	3.75 (.90)	3.73 (1.01)	3.95 (.73)	3.81 (.87)	4.09 (.61)	3.83 (.78)	3.97 (.88)	3.64 (.85)
Interpersonal Traits	3.85 (.66)	3.59 (.99)	3.86 (.80)	3.64 (1.01)	3.83 (.79)	3.43 (.94)	3.60 (.98)	3.58 (.62)
Overall Rating	3.26 (.72)	2.97 (.73)	3.45 (.71)	3.37 (.79)	3.38 (.64)	3.14 (.81)	3.50 (.82)	3.15 (.69)
Perceived Learning	2.57 (.99)	2.62 (1.39)	3.00 (.98)	3.03 (.90)	2.97 (1.05)	2.91 (1.00)	3.12 (.85)	3.09 (.92)
Actual Learning	3.30 (1.18)	3.80 (1.16)	3.80 (1.21)	3.52 (1.46)	3.28 (1.37)	3.42 (1.32)	3.64 (.99)	3.54 (1.56)

Table 3

Means and Standard Deviations of Teaching Evaluation Outcomes by Professor Gender

	Female Professor	Male Professor
Teaching Effectiveness	3.97* (.74)	3.78* (.85)
Interpersonal Traits	3.82* (.79)	3.54* (.92)
Overall Rating	3.40* (.69)	3.17* (.78)
Perceived Learning	2.93 (1.00)	2.92 (1.03)
Actual Learning	3.51 (1.26)	3.53 (1.37)

Note. The asterisks indicate a significant different within each row, comparing the female and male professor conditions ($p < .05$).

APPENDIX B

Link to Study 1 Materials:
https://purdue.ca1.qualtrics.com/jfe/form/SV_2h4a5VaMdrRsqJD

Link to Study 2 Materials:
https://purdue.ca1.qualtrics.com/jfe/form/SV_3fIqO6kpxOvsgkt

Link to Pilot Study Table
https://purdue.ca1.qualtrics.com/jfe/form/SV_3UfWWHWzDPqfQNv