Innov High Educ
DOI 10.1007/s10755-014-9313-4

What's in a Name: Exposing Gender Bias in Student Ratings of Teaching

Lillian MacNell · Adam Driscoll · Andrea N. Hunt

Abstract Student ratings of teaching play a significant role in career outcomes for higher education instructors. Although instructor gender has been shown to play an important role in influencing student ratings, the extent and nature of that role remains contested. While difficult to separate gender from teaching practices in person, it is possible to disguise an instructor's gender identity online. In our experiment, assistant instructors in an online class each operated under two different gender identities. Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender, demonstrating gender bias. Given the vital role that student ratings play in academic career trajectories, this finding warrants considerable attention.

Keywords gender inequality · gender bias · student ratings of teaching · student evaluations of instruction

Lillian MacNell is a doctoral candidate in Sociology at North Carolina State University. She received her Master's degree in Sociology at the University of Central Florida. Her research and teaching interests include food access, food justice, and the environment.

Adam Driscoll is Assistant Professor of Sociology at the University of Wisconsin-La Crosse. He received his Master's degree in Sociology at East Carolina University and his Ph.D. in Sociology at North Carolina State University. His research and teaching focus upon the environmental impacts of industrial agriculture and effective online pedagogy.

Andrea N. Hunt has a Ph.D. in Sociology from North Carolina State University and is currently Assistant Professor in Sociology and Family Studies at the University of North Alabama. Her research interests include gender, race and ethnicity, mentoring in undergraduate research, engaging teaching practices, and the role of academic advising in student retention.

L. MacNell (✉)
Department of Sociology and Anthropology, 334 1911 Building, Campus Box 8107,
Raleigh, North Carolina 27695, USA
e-mail: loconne@ncsu.edu

A. Driscoll
University of Wisconsin-La Crosse, La Crosse, WI, USA
e-mail: adriscoll@uw.lax.edu

A. N. Hunt
University of North Alabama, Florence, AL, USA
e-mail: ahunt3@una.edu

Published online: 05 December 2014

Student ratings of teaching are often used as an indicator of the quality of an instructor's teaching and play an important role in tenure and promotion decisions (Abrami, d'Apollonia, & Rosenfield, 2007; Benton & Cashin, 2014). Gender bias in these ratings constitutes an important form of inequality facing women in academia that is often unaccounted for in such decisions. Students perceive, evaluate, and treat female instructors quite differently than they do male instructors (Basow, 1995; Centra & Gaubatz, 2000; Feldman, 1992; Young, Rush, & Shaw, 2009). While a general consensus exists that gender plays a vital role in how students perceive and interact with their instructors, there is conflicting evidence as to whether or not this translates into a bias in student ratings due to variations in several mediating factors such as teaching styles and subject material.

Prior studies of student ratings of instruction have been limited in their ability to test for the existence of gender bias because it is difficult to separate the gender of an instructor from their teaching practices in a face-to-face classroom. In online courses, however, students usually base the categorization of their instructor's gender on the instructor's name and, if provided, photograph. It is possible for students to believe that their instructor is actually a man, based solely on a name or photograph, when in reality she is a woman, or vice versa. Therefore, the online environment affords researchers a unique opportunity to assign one instructor two different gender identities in order to understand whether or not differences in student ratings are a result of differences in teaching or simply based on unequal student expectations for male and female instructors. Such experimentation allows researchers to control for potentially confounding factors and therefore attribute observed differences solely to the variable of interest—in this case, the perceived gender of the instructor (Morgan & Winship, 2007).

This study analyzed differences in student ratings of their instructors[1] from an online course, independent of actual gender. The course professor randomly assigned students to one of six discussion groups, two of which the professor taught directly. The other four were taught by one of two assistant instructors—one male and one female. Each instructor was responsible for grading the work of students in their group and interacting with those students on course discussion boards. Each assistant instructor taught one of their groups under their own identity and the second group under the other assistant instructor's identity. Thus, of the two groups who believed they had the female assistant instructor, one actually had the male. Similarly, of the two groups who believed they had the male assistant instructor, one actually had the female (see Table 1). At the end of the course, the professor asked students to rate their instructor through the use of an online survey. This design created a controlled experiment that allowed us to isolate the effects of the gender identity of the assistant instructors, independent of their actual gender. If gender bias was present, than the students from the two groups who believed they had a female assistant instructor should have given their instructor significantly lower evaluations than the two groups who believed they had a male assistant instructor.

Student Ratings of Teaching

Though far from perfect, student ratings of teaching provide valuable feedback about an instructor's teaching effectiveness (Svinicki & McKeachie, 2010). They may be reliably interpreted as both a direct measure of student satisfaction with instruction and as an indirect

[1] To clarify the language we use throughout the paper, we refer to all three persons responsible for grading and directly interacting with students as "instructors." The course "professor" was the person responsible for course design and content preparation, while the two "assistant instructors" worked under the professor's direction to manage and teach their respective discussion groups.

Table 1 Experimental Design.

Discussion Group	Instructor's Perceived Gender	Instructor's Actual Gender
Group A (*n* =8)	Female	Female
Group B (*n* =12)	Female	Male
Group C (*n* =12)	Male	Female
Group D (*n* =11)	Male	Male

measure of student learning (Marsh, 2007; Murray, 2007). They also play an important role in the selection of teaching award winners, institutional reviews of programs, and student course selection (Benton & Cashin, 2014). More importantly to the careers of educators, these ratings are "used by faculty committees and administrators to make decisions about merit increases, promotion, and tenure" (Davis, 2009, p. 534). In particular, quantitative evaluations of instructors' overall teaching effectiveness are frequently emphasized in personnel decisions (Centra & Gaubatz, 2000). Given the widespread reliance on student ratings of teaching and their effect on career advancement, any potential bias in those ratings is a matter of great consequence.

Gender Bias in Academia

Sociological studies of gender and gender inequality are careful to distinguish between sex (a biological identity) and gender (a socially constructed category built around cultural expectations of male- and female-appropriate behavior). Gender is part of an ongoing performance based on producing a configuration of behaviors that are seen by others as normative. West and Zimmerman (1987) suggested that people engage in gendered behaviors not only to live up to normative standards, but also to minimize the risk of accountability or gender assessment from others. Thus, gender is a process that is accomplished at the interactional level and reinforced through the organization of social institutions such as academia (Lorber, 1994). Gender then contributes to a hierarchal system of power relations that is embedded within the interactional and institutional levels of society and shapes gendered expectations and experiences in the workplace (Risman, 2004).

An examination of gender bias in student ratings of teaching must be framed within the broader context of the pervasive devaluation of women, relative to men, that occurs in professional settings in the United States (Monroe, Ozyurt, Wrigley, & Alexander, 2008). In general, Western culture accords men an automatic credibility or competence that it does not extend to women (Johnson, 2006). Stereotypes that women are less logical, less confident, and occupy lower positions still pervade our organizational structures (Acker, 1990). Conversely, men are automatically assumed to have legitimate authority, while women must prove their expertise to earn the same level of respect. This disparity has been well documented in the field of academia, where men tend to be regarded as "professors" and women as "teachers" (Miller & Chamberlin, 2000) and women face a disparate amount of gender-based obstacles, relative to men (Morris, 2011).

In experiments where researchers gave students identical articles to evaluate—half of which bore a man's name and half of which bore a woman's—the students rated the research they thought had been done by men more highly (Goldberg, 1968; Paludi & Strayer, 1985). In a similar study, college students evaluated two hypothetical applicants for a faculty position and tended to judge the male candidate as more qualified despite the fact that both applicants had identical credentials (Burns-Glover & Veith, 1995). Additionally, a study of student evaluations of instructors' educational attainment revealed that students misattribute male instructors' education upward and female instructors' education downward (Miller & Chamberlin, 2000). Overall, women in academia tend to be regarded as less capable and less accomplished than men, regardless of their actual achievements and abilities.

Gender Role Expectations

Students often expect their male and female professors to behave in different ways or to respectively exhibit certain "masculine" and "feminine" traits. Commonly held masculine, or "effectiveness," traits include professionalism and objectivity; feminine, or "interpersonal," traits include warmth and accessibility. Students hold their instructors accountable to these gendered behaviors and are critical of instructors who violate these expectations (Bachen, McLoughlin, & Garcia, 1999; Chamberlin & Hickey, 2001; Dalmia, Giedeman, Klein, & Levenburg, 2005; Sprague & Massoni, 2005). Consequently, instructors who adhere to gendered expectations are viewed more favorably by their students (Andersen & Miller, 1997; Bennet, 1982). When female instructors exhibit strong interpersonal traits, they are viewed comparably to their male counterparts. When female instructors fail to meet these gendered expectations, however, they are sanctioned, while male instructors who do not exhibit strong interpersonal traits are not (Basow & Montgomery, 2005; Basow, Phelan, & Capotosto, 2006). At the same time, students are less tolerant of female instructors whom they perceive as lacking professionalism and objectivity than they are of male instructors who lack the same qualities (Bennet, 1982). In general, "students' perceptions and evaluations of female faculty are tied more closely to their gender expectations than for male faculty" (Bachen et al., 1999, p. 196).

These different standards can place female instructors in a difficult "double-bind," where gendered expectations (that women be nurturing and supportive) conflict with the professional expectations of a higher-education instructor (that they be authoritative and knowledgeable) (Sandler, 1991; Statham, Richardson, & Cook, 1991). On the one hand, students expect female instructors to embody gendered interpersonal traits by being more accessible and personable. However, these same traits can cause students to view female instructors as less competent or effective. On the other hand, female instructors who are authoritative and knowledgeable are violating students' gendered expectations, which can also result in student disapproval. Therefore, female instructors are expected to be more open and accessible to students *as well as* to maintain a high degree of professionalism and objectivity. Female instructors who fail to meet these higher expectations are viewed as less effective teachers than men (Basow, 1995).

Male instructors, however, are rated more highly when they exhibit interpersonal characteristics in addition to the expected effectiveness characteristics (Andersen & Miller, 1997). In other words, female instructors who fail to exhibit an ideal mix of traits are rated lower for not meeting expectations, while male instructors are not held to such a standard. Consequently, gendered expectations represent a greater burden for female than male instructors (Sandler, 1991; Sprague & Massoni, 2005). An important manifestation of that disparity is bias in student ratings of instructors, where female instructors may receive lower ratings than males, not because of differences in teaching but for failing to meet gendered expectations.

Methodological Concerns with Previous Studies of Gender Bias

Studies of gender bias in student ratings of instruction have presented complicated and sometimes contradictory results. Sometimes men received significantly higher ratings (Basow & Silberg, 1987; Sidanius & Crane, 1989), sometimes women (Bachen et al., 1999; Rowden & Carlson, 1996), and sometimes neither (Centra & Gaubatz, 2000; Feldman, 1993). The variety of results in these studies suggests that gender does play a role in students' ratings of their instructors, but that it is a complex and multifaceted one (Basow et al., 2006).

One reason why prior research on gender bias in student ratings of teaching has provided such inconclusive results may lie in the research design of these previous studies. A large portion of research on student ratings of teaching directly utilized those ratings for their data (e.g. Basow, 1995; Bennett, 1982; Centra, 2007; Centra & Gaubatz, 2000; Marsh, 2001). This strategy allows for the analysis of a large amount of data, but it does not control for differences in actual teaching and therefore may fail to capture gender bias in student ratings. Studies that compare student ratings of instructors explore whether or not there are differences—not whether or not those differences are the result of gender bias (Feldman, 1993). For example, a study of ratings may find that a female instructor received significantly lower scores than a male peer, but it could not assess whether that indicates a true difference in teaching quality. Perhaps she was not perceived as warm and engaging; failing to meet the gendered expectations of the students, she may have been rated more poorly than her male peer despite being an equally effective instructor. Similarly, the lack of a gender disparity in student ratings of instruction could actually obscure a gender bias if at a particular institution the female faculty members were, on average, stronger instructors than the males, yet were being penalized by the students due to bias (Feldman, 1993).

Additionally, a number of situational elements may serve to sway student ratings of male verses female instructors as male and female professors tend to occupy somewhat different teaching situations. Men are overrepresented in the higher ranks of academic positions as well as in STEM fields. They are also more likely to teach upper-level courses whereas women are more likely to teach introductory courses (Simeone, 1987; Statham et al., 1991). Women are also more likely than men to be employed in full-time non-tenure track positions as well as in part-time positions (Curtis, 2011). These factors are highly relevant because instructor rank, academic area, and class level of the course have all been found to directly impact student ratings of instruction (Feldman, 1993; Liu, 2012). All of these factors serve to complicate the relationship between instructor gender and student ratings of instruction and obfuscate the conclusions that can be drawn from direct studies of such ratings. Studies of actual student ratings of instruction may tell us more about women's position in academia than about actual gender bias in student ratings. In contrast, experimental studies allow the researcher to control for both the quality and character of the teaching as well as the academic position of the instructor; ensuring that any differences registered in student ratings indicate, as much as possible, a bias rather than an actual difference in teaching (Feldman, 1993).

Research Question and Related Hypotheses

The fundamental question examined in this study is whether or not students rate their instructors differently on the basis of what they perceive those instructors' gender to be. We expected that there would be no difference between the ratings for the actual male and female instructors in the course as every attempt was made to minimize any differences in interaction and teaching. However, we expected that student ratings of instructors would reflect the different expectations for male and female instructors discussed above. Instructors whom students perceived to be male would be afforded an automatic credibility on their competence and professionalism. Furthermore, they would not be penalized for any perceived deficiency in their interpersonal skills. Therefore, we expected that students would rate the instructors they *believed* to be male more highly than ones they believed to be female, regardless of the instructors' actual gender.

The Study and Methodology

This study examined gender bias in student ratings of teaching by falsifying the gender of assistant instructors in an online course and asking students to evaluate them along a number of instructional criteria. By using a 2-by-2 experimental design (see Table 1), we were able to compare student evaluations of a perceived gender while holding the instructor's actual gender (and any associated differences in teaching style) constant. Any observed differences in how students rated one perceived gender versus the other must have therefore derived from bias on the students' part, given that the exact same two instructors (one of each gender) were being evaluated in both cases.

Subjects

Data were collected from an online introductory-level anthropology/sociology course offered during a five-week summer session at a large (20,000+), public, 4-year university in North Carolina. The University's institutional review board had approved this study (IRB# 2640). The course fulfilled one of the university's general education requirements, and the students represented a range of majors and grade levels. The majority of the participants were traditional college-aged students with a median age of 21 years. The instructors taught the course entirely through a learning management system and students' only contact with their instructors was either through e-mail or comments posted on the learning management system. The professor delivered course content through assigned readings and written PowerPoint slideshow lectures. The course was broken up into nine different content sections. For each section, students were required to read the assigned material and make a series of posts on a structured discussion board. The course had 72 students who were randomly divided into six discussion groups for the entirety of the course. All discussion board activity took place within the assigned discussion group. Each discussion group had one instructor responsible for moderating the discussion boards and grading all assignments for that group. The course professor took two groups and divided the remaining four between the two assistant instructors, each taking one group under their own identity and a second under their fellow assistant instructor's identity (see Table 1). All instructors were aware of the study being conducted and cooperated fully.

The section discussion boards were the primary source of interaction between students and the course instructors and, as such, represented 30% of the students' final grades. The discussion boards were also an important part of student learning because they were the main arena in which students could analyze and voice questions about course concepts and material. The instructor assigned to each discussion group maintained an active presence on each discussion board, offering comments and posing questions. The instructor also graded students' posts and provided detailed feedback on where students had lost points. The two assistant instructors for the four discussion groups employed a wide range of strategies so as to maintain consistency in teaching style and grading. The two assistant instructors composed personal introduction posts that indicated similar biographical information and background credentials. They posted on the discussion boards and graded assignments at the same time of day three days each week to ensure that no group received significantly faster or slower feedback than others. The professor provided detailed grading rubrics for the discussion boards, and the instructors coordinated their grading to ensure that these rubrics were applied to students' work equitably.[2]

[2] A one-way ANOVA test confirmed that there was no significant variation among all six groups' discussion board grades and overall grades for the course.

Toward the end of the course the professor sent students reminder e-mails requesting that they complete an online evaluation of their instructor. These evaluations were explained as serving the purpose of providing the professor with feedback about the instructors' performance. The survey asked students to rate their instructor on various factors such as accessibility, effectiveness, and overall quality. Over 90% of the class completed the evaluation. For the purpose of this study, we only analyzed data from the discussion groups assigned to the assistant instructors, leaving us with 43 subjects.

Instrument

The instructor evaluation consisted of 15 closed-ended questions that ask students to rate their instructors on a variety of measures using a five-point Likert scale (1 = Strongly disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, 5 = Strongly agree). The survey had six questions designed to measure effectiveness traits (e.g. professionalism, knowledge, and objectivity) and six questions designed to measure interpersonal traits (e.g. respect, enthusiasm, and warmth). In addition, there were two questions designed to measure communication skills and one question that asked students to evaluate the instructor's overall quality as a teacher. We also asked students to indicate which discussion group they were in and to provide basic demographic and academic background information including gender, age, year in school, and number of credit hours currently being taken. All students fully completed the evaluation, leaving us with no missing data.

We performed all analyses with the 13th version of the Stata statistical analysis program. We used exploratory factor analysis to test how well the separate questions reflected a common underlying dimension. Principal component factor analysis revealed that 12 of our items characterized a single factor for which the individual factor loadings ranged from .7370 to .9489; sufficiently high to justify merging them into a single index (Hair, Anderson, Tatham, & Black, 1998). This indicates that those 12 questions on our survey were all measuring the same latent variable, which we interpret to be a general evaluation of the instructor's teaching. A reliability test yielded a Cronbach's alpha above .950 for the 12 questions. In order to confirm the factor structure, we used structural equation modeling to test a single latent variable indicated by our 12 separate questions. Our model was a strong fit to the data (N =43, $\chi^2(47)$ =59.18 (not significant), RMSEA =0.078, CFI =0.980, SRMR =0.043) with all loadings significant at the p <0.001 level. Therefore, we extracted a factor score, *student ratings index*, which weighed each question by how strongly it loaded onto the single factor, providing us with a single representation of how well each student evaluated their instructor's teaching.

Analysis

To test for the existence of gender bias in student ratings of teaching, we made two types of comparisons. First we compared across the *actual* gender of the assistant instructor, combining the two groups that had the female assistant instructor (one of which thought they had a male) into one category and doing the same with the two groups that had the male assistant instructor. Second, we compared across the *perceived* gender of the assistant instructor, combining the two groups that thought they had a female assistant instructor (one of which was actually a male) into one category and doing the same with the two groups that thought they had a male assistant instructor. We made both comparisons for the 12 individual questions, as well as the *student ratings index*. We used Welch's t-tests (an adaptation of the Student's t-test that does not assume equal variance) to establish the statistical significance of each difference. We also ran two general linear multivariate analyses of variance (MANOVAs) on the set of 12 variables to test the effects of instructor gender (perceived and actual) on all of the questions considered as a group. A MANOVA allows a researcher to test a set of correlated dependent variables and conduct a single, overall comparison between the groups formed by categorical independent variables (Garson, 2012). This F-test of all means addresses the potential for false positive findings as the result of multiple comparisons.[3]

Results

Student Ratings of Perceived and Actual Gender

By comparing differences across the *actual* gender of the assistant instructor with those observed across the *perceived* gender of the instructor it is possible to observe whether or not students rated their instructors differently depending on the gender of the instructor. The results of this comparison are found in Table 2.

Our MANOVAs indicate that there is a significant difference in how students rated the perceived male and female instructors ($p < 0.05$), but not the actual male and female instructors. When looking at the individual questions as well as the *student ratings index*, there are no significant differences between the ratings of the actual male and female instructor (the first and second columns in Table 2). Students in the two groups that had the female assistant instructor (one of which thought they had a male) did not rate their instructor any differently than did the students in the two groups that had the male assistant instructor. The left two columns of Fig. 1 provide a graphic representation of this comparison for the *student ratings index*. The overlapping error bars (+/- one standard error) indicate the lack of a significant difference between how students rated the actual male and female assistant instructors.

When comparing between the perceived gender identities of the instructors (the fourth and fifth columns in Table 2), we found that the male identity received significantly higher scores on professionalism, promptness, fairness, respectfulness, enthusiasm, giving praise, and the *student ratings index*.[4] Looking at the R-squares, all seven of these comparisons yielded a medium sized effect. It is worth noting, particularly given the small sample size, that the male instructor identity also received higher scores on the other six questions, though not to a statistically significant degree. Students in the two groups that perceived their assistant instructor to be male rated their instructor significantly higher than did the students in the two groups that perceived their assistant instructor to be female, regardless of the actual gender of the assistant instructor. This comparison is represented graphically by the right two columns of Fig. 1, where a clear difference can be observed.

[3] We acknowledge that the application of parametric analytical techniques (ANOVA, MANOVA, and t-tests) to ordinal data (the Likert scale responses) remains controversial among social scientists and statisticians. (See Knapp (1990) for a relatively balanced review of the debate.) We side with the arguments of Gaito (1980) and Armstrong (1981) and argue that it is appropriate to do so in our case as the concept being measured is interval, even if the data labels are not. This practice is common within higher education research. (e.g. Centra & Gaubatz [2000] Young, Rush, & Shaw [2009]; Basow [1995]; and Knol et al. [2013])

[4] While we acknowledge that a significance level of .05 is conventional in social science and higher education research, we side with Skipper, Guenther, and Nass (1967), Labovitz (1968), and Lai (1973) in pointing out the arbitrary nature of conventional significance levels. Considering our study design, we have used a significance level of .10 for some tests where: 1) the results support the hypothesis and we are consequently more willing to reject the null hypothesis of no difference; 2) our hypothesis is strongly supported theoretically and by empirical results in other studies that use lower significance levels; 3) our small n may be obscuring large differences; and 4) the gravity of an increased risk of Type I error is diminished in light of the benefit of decreasing the risk of a Type II error (Labovitz, 1968; Lai, 1973).

Table 2 Comparison of means of student ratings of teaching across the actual gender of the assistant instructor and the perceived gender of the assistant instructor

Question	Actual Female	Actual Male	Difference	Perceived Female	Perceived Male	Difference
Caring	4.00 (1.257)	3.87 (0.868)	0.13 (*0.004*)	3.65 (1.226)	4.17 (0.834)	-0.52 (*0.071*)
Consistent	3.80 (1.322)	3.70 (1.020)	0.10 (*0.002*)	3.50 (1.357)	3.96 (0.928)	-0.47 (*0.045*)
Enthusiastic	4.05 (1.191)	3.78 (0.850)	0.27 (*0.019*)	3.60 (1.314)	4.17 (0.576)	-0.57† (*0.112*)
Fair	4.05 (1.050)	3.78 (0.951)	0.27 (*0.018*)	3.50 (1.192)	4.26 (0.619)	-0.76* (*0.188*)
Feedback	4.10 (1.252)	3.83 (1.029)	0.27 (*0.015*)	3.70 (1.380)	4.17 (0.834)	-0.47 (*0.054*)
Helpful	3.65 (1.309)	3.83 (0.834)	-0.18 (*0.008*)	3.50 (1.192)	3.96 (0.928)	-0.46 (*0.049*)
Knowledgeable	4.20 (1.056)	4.09 (0.949)	0.11 (*0.003*)	3.95 (1.191)	4.30 (0.765)	-0.35 (*0.038*)
Praise	4.35 (0.988)	4.09 (0.900)	0.26 (*0.020*)	3.85 (1.089)	4.52 (0.665)	-0.67* (*0.153*)
Professional	4.30 (1.218)	4.35 (0.935)	-0.05 (*0.000*)	4.00 (1.414)	4.61 (0.499)	-0.61† (*0.124*)
Prompt	4.10 (1.252)	3.87 (0.919)	0.23 (*0.013*)	3.55 (1.356)	4.35 (0.573)	-0.80* (*0.191*)
Respectful	4.30 (1.218)	4.35 (0.935)	-0.05 (*0.001*)	4.00 (1.414)	4.61 (0.499)	-0.61† (*0.124*)
Responsive	4.00 (1.124)	3.57 (0.843)	0.43 (*0.052*)	3.65 (1.137)	3.87 (0.869)	-0.22 (*0.013*)
Student Rating Index	0.09 (1.165)	-0.08 (0.850)	0.17 (*0.008*)	-0.33 (1.267)	0.284 (0.584)	-0.61† (*0.128*)
N	20	23		20	23	

Note: Each cell contains the mean student response for the question with the standard deviations in parentheses. The cells in the Difference columns contain the difference between the means with the *r*-squared in italics and parentheses. Welch's *t*-tests were used to establish the significance of the observed differences.

† p < =0.10.
* p < =0.05.

Discussion

With the design of this experiment, we are able to attribute any differences between how students rated the two perceived genders to gender bias as the students actually evaluated the same two instructors in each case. Our findings support the existence of gender bias in that

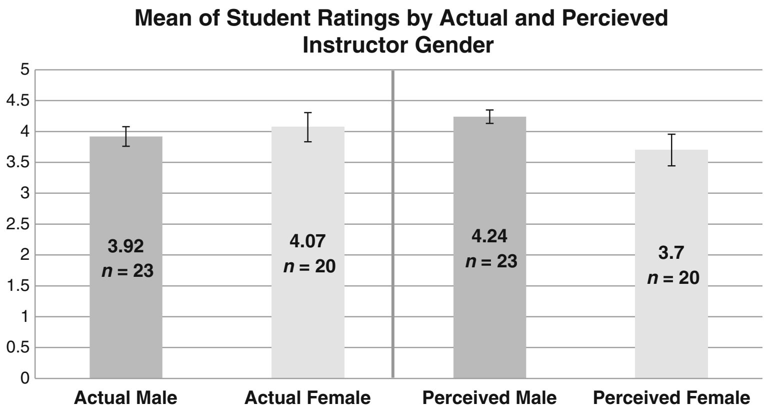

Figure 1 Comparison of the mean of student ratings across actual instructor gender (left two columns) and perceived instructor gender (right two columns). The difference between the right two columns is significant to the p<=0.10 level.

students rated the instructors they perceived to be female lower than those they perceived to be male, regardless of teaching quality or actual gender of the instructor. The perceived female instructor received significantly lower ratings on six of the 12 metrics on the survey, as well as on the *student ratings index*.

The difference between how students rated the two perceived genders stands in stark contrast to the fact that neither the actual male nor actual female instructor received significantly higher ratings than the other. Both instructors performed equally well from the students' perspective. However, in both cases the *same* instructor received different ratings depending solely on their perceived gender. In other words, when the actual male instructor was perceived to be female, he received significantly lower ratings than when he was perceived to be a male. For example, when the actual male and female instructors posted grades after two days *as a male*, this was considered by students to be a 4.35 out of 5 level of promptness, but when the same two instructors posted grades at the same time *as a female*, it was considered to be a 3.55 out of 5 level of promptness. In each case, the same instructor, grading under two different identities, was given lower ratings half the time with the only difference being the perceived gender of the instructor. Similarly, students rated the perceived female instructors an average of 0.75 points lower on the question regarding fairness, despite both instructors utilizing the same grading rubrics and there being no significant differences in the average grades of any of the groups. These findings support the argument that male instructors are often afforded an automatic credibility in terms of their professionalism, expertise, and effectiveness as instructors. Despite the fact that the students were equally satisfied with the promptness and fairness of the *actual* instructors, the instructor that students perceived to be male was considered to be more effective.

Similarly, both actual instructors demonstrated the same level of interpersonal interaction in their attempts to create a sense of immediacy in the online classroom. Yet the perceived male instructor received higher ratings on all six interpersonal measures, three of them significantly. We contend that female instructors are *expected* to exhibit such traits and therefore are not rewarded when they do so, while male instructors are perceived as going above and beyond expectations when they exhibit these traits. In other words, students have higher interpersonal standards for their female instructors (Sandler, 1991). Our findings support the existence of this bias. In the online environment, it is more difficult to create immediacy through verbal communication, and nonverbal communication and body language are eliminated entirely (O'Sullivan, Hunt, & Lippert, 2004). Students sanctioned the perceived female instructor for failing to demonstrate strong interpersonal traits, yet did not do the same for the perceived male instructor. Both instructors were working within the same confines of online, text-based communications, but students only penalized the instructor they perceived to be female for this shortcoming.

Although this experiment was conducted in the online environment, we believe that the findings apply more broadly to all student ratings of teaching. Rather than testing for gender bias in the online environment, we used this environment as a natural laboratory to test for the existence of gender bias in student ratings as a whole. We argue that the demonstrated bias exists in the general student population and will manifest itself in both online and face-to-face classrooms. The combination of higher expectations and lower automatic credibility translates into very real differences in student ratings of female versus male instructors. Though it is easier to affect interpersonal characteristics in a face-to-face environment, the fact remains that some professors are *expected* to do so while others are given a ratings boost for those same behaviors.

Because student ratings of teaching are considered an important measure of teaching proficiency, the existence of gender bias in those scores needs to be better understood and acknowledged within the institutional framework of our higher-education system. These results provide strong evidence that gender bias exists in student ratings of their instructors, but more work is needed. First and foremost, these results need to be replicated in other similar online classes. A single case study cannot establish a broad pattern. However, it does suggest the existence of one and provides incentive for further exploration. Additional studies of this type could lend weight to these findings and better establish the existence of this bias throughout academia. Additionally, courses in other subject areas with a variety of both male and female instructors should follow a similar model to corroborate these findings.

Conclusions

Our findings show that the bias we saw here is *not* a result of gendered behavior on the part of the instructors, but of actual bias on the part of the students. Regardless of actual gender or performance, students rated the perceived female instructor significantly more harshly than the perceived male instructor, which suggests that a female instructor would have to work harder than a male to receive comparable ratings. If female professors and instructors are continually receiving lower evaluations from their students for no other reason than that they are women, then this particular form of inequality needs to be taken into consideration as women apply for academic jobs and come up for promotion and review.

These findings represent an important contribution to existing debates over the validity of student ratings of teaching. (See Benton & Cashin, 2014; Perry & Smart, 2007; and Theall, Abrami, & Mets, 2001 for reviews.) These debates have highlighted a number of weaknesses and shortcomings of student ratings of teaching as a reflection of the quality of instruction being rated (Greenwald, 1997; Johnson, 2003; Svanum & Aigner, 2011). They have also shown that there is substantial room for updating and improving how student ratings of teaching are collected, interpreted, and utilized (Hampton & Reiser, 2004; Subramanya, 2014). However, for better or worse, they remain one of the primary tools used to evaluate educators' teaching for the purposes of promotion and tenure decisions (Davis, 2009; Svinicki & McKeachie, 2010). This study demonstrates that gender bias is an important deficiency of student ratings of teaching. Therefore, the continued use of student ratings of teaching as a primary means of assessing the quality of an instructor's teaching systematically disadvantages women in academia. As this limitation is one of numerous problems associated with the emphasis on quantitative student ratings of teaching, 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 (Hampton & Reiser, 2004; Morrison & Johnson, 2013).

It is also worth noting that this experiment is only scratching the surface of what is possible with gender studies in the online environment. The online environment presents a unique opportunity to experiment directly with gender identity. Analyzing the difference in online behavior of individuals when they perceive that they are interacting with a male or female could provide a wealth of data on how gender is constructed and treated. We hope that this experiment serves as a model for future work that will enhance our ability to test for gender bias in order to further our understanding of its basis, means of perpetuation, and potential avenues of amelioration.

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. 385–445). Dordrecht, The Netherlands: Springer.

Acker, J. (1990). Hierarchies, job, and bodies: A theory of gendered organizations. *Gender and Society, 4*, 81–95.

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

Armstrong, G. D. (1981). Parametric statistics and ordinal data: A pervasive misconception. *Nursing Research, 30*, 60–62.

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

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

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-rating of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*, 91–106.

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*, 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*, 308–314.

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.

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

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*, 69–80.

Centra, J. A. (2007). *Differences in responses to the student instructional report: Is it bias?* Princeton, NJ: Educational Testing Service.

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

Chamberlin, M. S., & Hickey, J. S. (2001). Student evaluations of faculty performance: The role of gender expectations in differential evaluations. *Educational Research Quarterly, 25*, 3–14.

Curtis, J. W. (2011). *Persistent inequity: Gender and academic employment.* Report from the American Association of University Professors. Retrieved from http://www.aaup.org/NR/rdonlyres/08E023AB-E6D8-4DBD-99A0-24E5EB73A760/0/persistent_inequity.pdf

Dalmia, S., Giedeman, D. C., Klein, H. A., & Levenburg, N. M. (2005). Women in academia: An analysis of their expectations, performance and pay. *Forum on Public Policy, 1*, 160–177.

Davis, B. G. (2009). *Tools for teaching* (2nd ed.). San Francisco, CA: Jossey-Bass.

Feldman, K. A. (1992). College students' views of male and female college teachers: Evidence from the social laboratory and experiments – Part 1. *Research in Higher Education, 33*, 317–375.

Feldman, K. A. (1993). College students' views of male and female college teachers: Evidence from the social laboratory and experiments – Part 2. *Research in Higher Education, 34*, 151–211.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564–567.

Garson, G. D. (2012). *General linear models: Multivariate GLM & MANOVA/MANCOVA*. Asheboro, NC: Statistical Associates.

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

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*, 1182–1186.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis with readings* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hampton, S. E., & Reiser, R. A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education, 45*, 497–527.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York, NY: Springer.

Johnson, A. (2006). *Power, privilege, and difference*. Boston, MA: McGraw-Hill.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research, 39*, 121–123.

Knol, M. H., Veld, R., Vorst, H. C. M., van Driel, J. H., & Mellenbergh, G. J. (2013). Experimental effects of student evaluations coupled with collaborative consultation on college professors' instructional skills. *Research in Higher Education, 54*, 825–850.

Labovitz, S. (1968). Criteria for selecting a significance level: A note on the sacredness of.05. *The American Sociologist, 3*, 220–222.

Lai, M.K. (1973). *The case against tests of statistical significance*. Report from the Teacher Education Division Publication Series. Retrieved from http://files.eric.ed.gov/fulltext/ED093926.pdf

Liu, O. L. (2012). Student evaluation of instruction: In the new paradigm of distance education. *Research in Higher Education, 53*, 471–486.

Lorber, J. (1994). *Paradoxes of gender*. New Haven, CT: Yale University Press.

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal, 38*, 183–212.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, The Netherlands: Springer.

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

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*, 215–233.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, MA: Cambridge University Press.

Morris, L. V. (2011). Women in higher education: Access, success, and the future. *Innovative Higher Education, 36*, 145–147.

Morrison, K., & Johnson, T. (2013). Editorial. *Educational Research and Evaluation, 19*, 579–584.

Murray, H. G. (2007). Low-inference teaching behaviors and college teaching effectiveness: Recent developments and controversies. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 145–183). Dordrecht, The Netherlands: Springer.

O'Sullivan, P. D., Hunt, S. K., & Lippert, L. R. (2004). Mediated immediacy: A language of affiliation in a technological age. *Journal of Language and Social Psychology, 23*, 464–490.

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*, 353–361.

Perry, R. P., & Smart, J. C. (Eds.). (2007). *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Dordrecht, The Netherlands: Springer.

Risman, B. J. (2004). Gender as a social structure: Theory wrestling with activism. *Gender & Society, 18*, 429–450.

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

Sandler, B. R. (1991). Women faculty at work in the classroom, or, why it still hurts to be a woman in labor. *Communication Education, 40*, 6–15.

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

Simeone, A. (1987). *Academic women: Working toward equality.* South Hadley, MA: Bergin and Garvey.

Skipper, J. K., Guenther, A. C., & Nass, G. (1967). The sacredness of.05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist, 1*, 16–18.

Sprague, J., & Massoni, K. (2005). Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles, 53*, 779–793.

Statham, A., Richardson, L., & Cook, J. A. (1991). *Gender and university teaching: A negotiated difference*. Albany, NY: State University of New York Press.

Subramanya, S. R. (2014). Toward a more effective and useful end-of-course evaluation scheme. *Journal of Research in Innovative Teaching, 7*, 143–157.

Svanum, S., & Aigner, C. (2011). The influences of course effort, mastery and performance goals, grade expectancies, and earned course grades on student ratings of course satisfaction. *British Journal of Educational Psychology, 81*, 667–679.

Svinicki, M., & McKeachie, W. J. (2010). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (13th ed.). Belmont, CA: Wadsworth.

Theall, M., Abrami, P. C., & Mets, L. A. (Eds.). (2001). *The student ratings debate: Are they valid? How can we best use them?* San Francisco, CA: Jossey-Bass.

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

Young, S., Rush, L., & Shaw, D. (2009). Evaluating gender bias in ratings of university instructors' teaching effectiveness. *International Journal of Scholarship of Teaching and Learning, 3*, 1–14.