THE EFFECTS OF PROFESSORS' RACE AND GENDER ON STUDENT EVALUATIONS AND PERFORMANCE

SUSAN A., BASOW
Psychology Department, Lafayette College

STEPHANIE CODOS
Counseling Psychology, Lehigh Univeristy

JULIE L. MARTIN
Social Psychology, Duke Univeristy

This experimental study examined the effects of professor gender, professor race, and student gender on student ratings of teaching effectiveness and amount learned. After watching a three-minute engineering lecture presented by a computer-animated professor who varied by gender and race (African American, White), female and male undergraduates ($N = 325$) completed a 26-question student evaluation form and a 10-question true/false quiz on the lecture content. Contrary to predictions, male students gave significantly higher ratings than female students on most teaching factors and African American professors were rated higher than White professors on their hypothetical interactions with students. Quiz results, however, supported predictions: higher scores were obtained by students who had a White professor compared to those who had an African American professor, and by students who had a male professor compared to those who had a female professor. These results may be due to students paying more attention to the more normative professor. Thus, performance measures may be a more sensitive indication of race and gender biases than student ratings. The limited relationship between student ratings and student learning suggests caution in using the former to assess the latter.

Student ratings of professors play an extremely influential role in personnel decisions related to hiring, promotion, tenure, and salaries in the U.S. (Cashin, 1999). In a survey completed in 1998 by 4-year liberal arts colleges, 88.1% of schools reported "always using" student evaluations as a primary method to evaluate teacher performance, a percentage that has continued to increase in recent years (Seldin, 1999). Despite the important role that student ratings of professors play in employment decisions, there still is tremendous uncertainty about what exactly these ratings assess and whether they could reflect student gender and racial biases (Addison, Best, & Warrington, 2006; Basow & Martin, 2012; Reid, 2010; Sprague & Massoni, 2005). The current study examined the effects of professor gender and race on student ratings and performance using an experimental design.

Social psychological research has documented how a rater's perception of and reaction to another person can be affected by bias, either consciously or unconsciously (Biernat, 2003; Eagly & Karau, 2002; Phelan, Moss-Racusin, & Rudman, 2008.) In particular, gender and race stereotypes may create different expectations for different individuals. For example, because women are expected to be nurturant and caring, a woman's interpersonal

skills may be viewed more critically in a rater's overall evaluation than is the case when rating her male counterpart. Furthermore, women and minorities often must work harder to be perceived as equally competent as White men (the normative group), and it is far easier for them to "fall from grace" as well (Biernat, Fuegen, & Kobrynowicz, 2010; Foschi, 2000). That is, a mistake or negative quality is viewed more negatively when the person being rated is from a minority group than when the person is from the majority/ normative group. Thus, the same behavior, such as grading harshly, might be perceived more negatively by students if the professor is a woman or African American or Hispanic (who "should" be "nice" and "caring") than if the professor is a White man (who has greater legitimacy due to both race and gender). Such indeed seems to be the case. For example, Sinclair and Kunda (2000) found that students responded differentially to negative feedback (low grades) from a professor depending upon whether the professor was a man or a woman. When students received negative feedback from a woman, they rated her as less competent than they did a man.

In the student evaluation literature, there is more research on the effects of professor gender than professor race on student ratings, probably because non-White professors are such a minority in the professorate. Field research regarding potential biases in student ratings is challenging for other reasons as well. Professors in real life vary on many factors, all at the same time: gender, race, age, attractiveness, discipline, years of experience, personal style, etc. Thus, partialling out the effects of gender and/or race alone often is not possible. Laboratory research allows for a more controlled examination of individual variables and was used in the current study. Although the external validity of laboratory studies can be questioned, research suggests that they can generalize to the field. For example, Ambady and Rosenthal (1993) found that laboratory ratings of a brief (30 seconds) nonverbal video clip of a professor correlated significantly with the usual end-of-semester student ratings.

Previous research on the effects of professor gender on student ratings typically find few main effects, especially on global ratings of teacher effectiveness (e.g., Bennett, 1982; Feldman, 1993.) When effects are found, their size is small (e.g., Basow, 1995; Centra & Gaubatz, 2000.) Nonetheless, professor gender may affect student ratings in subtle ways, often in interaction with student gender or gender-typing of the discipline, and often only on certain aspects of teaching. In particular, although male faculty tend to be rated similarly by their male and female students, female faculty tend to be rated lower by their male students and sometimes higher by their female students (Basow, 1995; Basow & Silberg, 1987; Centra & Gaubatz, 2000; Feldman, 1993). Thus, unless student gender is examined in interaction with professor gender, the average ratings of male and female professors look similar. Male students are 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 have had (Basow, 2000; Basow, Phelan, & Capotosto, 2006). In contrast, female students often do choose women faculty as "best" and rate them higher than male faculty, especially on qualities related to "fairness" and "providing a comfortable classroom environment." These gendered teacher qualities also were found in Sprague and Massoni's (2005) study examining the descriptions of a student's best teacher. Best male teachers were more likely to be described as funny while best female teachers were more likely to be categorized as nurturing and caring by both female and male students. In general, a variety of studies have found that men faculty often are rated higher than women on questions related to scholarship/knowledge and dynamism/enthusiasm,

while women faculty often are rated higher than men on questions relating to faculty-student interactions (Bachen, McLoughlin, & Garcia, 1999; Basow & Montgomery, 2005; Bennett, 1982; Centra & Gaubatz, 2000).

There have been very few studies of the effects of professor race on student ratings, but those that have been conducted often find that African American and Hispanic faculty receive lower evaluations than White and Asian faculty (Basow & Martin, 2012). For example, using student evaluations of faculty at the top 25 liberal arts colleges in the U.S. posted on the website ratemyprofessor.com, Reid (2010) found that Black faculty, especially Black men, were evaluated more critically and given lower ratings on quality, helpfulness, and clarity than their White counterparts. (The main effect of professor race was qualified by an interaction with professor gender.) Although ratings on the website ratemyprofessor.com are not randomly collected, there is evidence that such ratings correlate significantly with actual on-site course ratings (e.g., Timmerman, 2008). Similar to Reid's results, Smith (2007) found that White faculty were rated consistently higher than Black faculty on global measures of overall teaching at a large university in the south U.S. Because these are naturalistic studies, it is not clear whether such differential ratings are due to bias, to actual differences in teaching effectiveness, or to other factors, such as subject matter being taught or teaching style. Some indication that bias may be involved comes from a study by Ho, Thomsen, and Sidanius (2009) who examined how ratings of professor intellectual competence and sensitivity to students related to ratings of overall teaching effectiveness in a sample of 5,655 randomly-selected students. Although the overall performance ratings of African American faculty did not differ from those of White faculty, students' perceptions of intellectual competence were a bigger factor in the overall performance evaluations of Black compared to White faculty. That is, there was a stronger correlation between intellectual competence ratings and overall performance ratings for the African American faculty than for the Caucasian faculty. This finding is consistent with social psychological research findings that lower status groups (e.g., women, African Americans) must "prove" competence when evaluated for high status positions (e.g., Foschi, 2000).

In one of the few laboratory studies on the effects of professor race (White, Asian, or African American) and gender on student evaluations, Bavishi, Madera, and Hebl (2010) asked entering college students to rate a Curricula Vita which varied by gender (indicated by name and title) and race/ethnicity (indicated by name and organizational memberships) as to the competency, legitimacy, and interpersonal skills of the hypothetical professor. Results revealed that African American professors, especially women, were rated the lowest on all three dimensions and Asian professors were rated lower than the White professors on interpersonal skills.

Another issue of concern in the student rating literature is the degree to which such ratings reflect actual teaching effectiveness rather than more subjective variables, such as bias or "liking" (see Basow & Martin, 2012). Previous research is mixed on the relationship between student ratings and student achievement: although student ratings are significantly correlated with student grades in university settings (as would be expected if students learn more from more highly-rated professors), the correlation is actually higher between expected grade and student ratings than between actual grade and student rating (Ducette & Kenney, 1982; Greenwald & Gillmore, 1997; Millea & Grimes, 2002). A recent meta-analysis (Clayson, 2009) found no published findings after 1990 that contained a statistically significant positive association between student learning and student ratings.

The effects of professor race and gender

on student learning have rarely been examined. In one laboratory study on the effects of professor gender and teaching style (expressive/nonexpressive) on student ratings and achievement, Basow (1990) found that professors (videotaped actors giving a short history lecture) in the expressive condition were rated more positively than the same professors in the nonexpressive condition, but these student ratings were not significantly correlated with student recall of lecture content based on a multiple-choice test. Instead, students who had the male lecturer scored the highest, especially when he was nonexpressive, perhaps because they found the more normative (i.e., male) professor more credible.

In order to increase understanding of factors impacting student learning in post-secondary environments, more experimental research is needed on the effects of professor race, professor gender, and student gender on learning outcomes and professor evaluations. The present study, the first of its kind, examined the effects of professor race (White, African American) and gender along with student gender on student ratings and performance after hearing a computer-based lecture given by a computer-animated professor. The brief lecture was on an engineering-related topic ("flow visualization") because previous research has shown the most negative attitudes from both male and female students toward women professors may occur in natural science and engineering fields (Basow, 1995; Basow & Montgomery, 2005). Our student evaluation measure was multi-factorial (tapping factors of Scholarship, Organization/Clarity, Instructor-Group Interaction, Instructor-Individual Student Interaction, and Dynamism/Enthusiasm), as professor race and gender may affect only certain aspects of teaching (Basow, 1998). Finally, we included a measure of performance (true/false test based on lecture content) to examine the relationship between student ratings of teaching and actual student learning.

In this 2 (professor gender) x 2 (professor race) x 2 (student gender) between-participants experimental study, we provided four groups of respondents with an identical 3-minute video clip adjusted to look and sound either male or female and either White or African American. We hypothesized the following: Hypothesis 1: Based on previous research (Bachen et al.; Basow, 1998; Centra & Gaubatz, 2000), there should be a significant interaction between professor gender and student gender such that male students should give lower ratings to the female professors than did female students on questions related to Scholarship and Dynamism/Enthusiasm. However, both male and female students should rate the women professors higher than the male professor on the interpersonal questions. Hypothesis 2: there should be a main effect of professor race such that White professors should receive higher ratings than African American professors (based on Bavishi et al., 2010; Ho et al., 2009; Reid, 2010; Smith, 2007). It is possible that professor race and professor gender would interact such that African American women professors received the lowest ratings (Bavishi et al., 2010). Hypothesis 3 was based on the effects of the independent variables on test performance and was exploratory. Although field research suggests student ratings should be modestly correlated with student achievement (Ducette & Kenney, 1982; Greenwald & Gillmore, 1997; Millea & Grimes, 2002), laboratory research (Basow, 1990; Basow & Distenfeld, 1985) has not found this to be the case. Instead, student performance may vary with the gender and/or race of the professor. If a White male professor is viewed as the most normative and credible, especially for a technical lecture, students may pay more attention and therefore score higher on the recall test.

Method

Participants

Three hundred and twenty-nine traditional-aged (18-22 years old) undergraduates (126 men, 203 women) from a small private liberal arts college in northeastern U.S. participated in this study for extra credit (from psychology, economics, or mathematics courses). The sample was 80.5% White/Caucasian (N = 265), 8.5% Asian/Pacific Islander (N = 28), 4.6% Hispanic (N = 15), 3.6% Black/African American (N = 12), 6% Middle Eastern/Arabic (N = 2), and 2.1% Other (N = 7). These demographics are proportional to those of the college (e.g., 80% White/Caucasian). The majority of students were either Social Sciences or Natural Sciences majors (N=119 and N=113, respectively). Grade Point Averages (GPAs) ranged from 2.0 to 4.0 (highest), with a mean of 3.32 and a mode of 3.00. Sixty-one participants were freshman, 128 were sophomores, 83 were juniors and 57 were seniors. Chi-Square analyses revealed that participants in the four experimental professor conditions (White Female, White Male, Black Female, Black Male) did not significantly ($p > .05$) vary by major, class year, or race/ethnicity (White vs. non-White). The GPAs of students in the four conditions also did not significantly differ ($p > .05$). Due to incomplete information for some variables used as covariates, the final sample consisted of 300 participants, 116 men and 184 women.

Materials

Professor manipulation. To control for attractiveness of the professor, a pilot study was conducted in which 10 college students (five male and five female) rated the attractiveness of 20 faces (five Black men, five Black women, five White men, and five White women obtained from an on-line photo database) on a 5-point Likert scale where *1* = "very unattractive" and *5* = "very attractive." The four final faces selected were chosen by first eliminating the faces from each race and gender that had the highest and the lowest mean attractiveness rating. The four faces that were most similarly rated (Ms = 3.1 – 3.3) were chosen; there were no significant differences among the four faces in average attractiveness ratings using the pilot data. However, using the entire sample, ratings of attractiveness did vary significantly for the male and female faces, $F(1, 321) = 10.62$, $p = .001$, $\eta 2 = .032$, with the female professors rated as more attractive ($M = 2.81$, $SD = .97$) than the male professors ($M = 2.45$, $SD = .99$).

CrazyTalk 6, a facial animation creator, was used to generate the animated professor for the four different conditions. Through the program, Black male, Black female, White male, and White female animated professors were created and brought to life through the use of the program's text-to-speech and facial animation features. The animation showed only a talking head in which the eyes and mouth moved.

Prior to conducting the study, a ten-minute sample lecture was obtained from an engineering professor at the college, but only the beginning three minutes were used. This lecture was presented by the animated professor in each of the four different conditions. To import the lecture into the program, a text-to-speech feature was used. The lecture was recorded using a female voice, and a voice changer was utilized for the male videos as a way to keep the tone and inflection of each voice constant. Computers were used to show the three-minute engineering lecture to each subject. The lecture, titled "The Art and Science of Flow Visualization," covered basic engineering concepts concerning the history and study of flow visualization. Although the lecture was very brief, Ambady and Rosenthal (1993) have shown that evaluations of teachers that were made after viewing a 30-second nonverbal video clip predicted end-of-the-semester evaluations.

Teacher evaluations. The main dependent measure was an adapted version of Hildebrand and Wilson's (1970) teacher evaluation form used in previous research (Basow, 1990; Basow & Montgomery, 2005; Basow & Silberg, 1987). The form included 25 questions that measured five factors with five questions each: Scholarship (Cronbach's $\alpha = .68$ in the current study), Organization/clarity ($\alpha = .80$), Instructor-group interaction ($\alpha = .84$), Instructor-individual interaction ($\alpha = .87$), and Dynamism/Enthusiasm ($\alpha = .86$), as well as a final question that measured the professor on a global measure. Each question on the form was rated using a 5-point Likert scale, where *1* = "well below average" and *5* = "well above average." The questions on the form examined specific components of the teacher's nature and ability. The final question on the rating form looked at a global evaluation of the teacher's overall ability, asking for a general rating of teaching ability. Given the nature of the computer program and its inability to gauge interaction or discussion, all questions pertaining to instructor-group interaction and instructor-individual interaction were adapted to contain hypothetical questions about the instructor's abilities to engage with his or her students. Therefore, students were asked how well they expected the professor to perform in such situations where he or she was engaging with students. Similar hypothetical questions have been used effectively in previous research (e.g., Bavishi et al., 2010).

Student achievement. The students were given a 10-question true/false-choice quiz created by the authors to measure their comprehension of the material. For half of the questions, the correct answer was True; for the other half, the correct answer was False. Thus, a chance score would be 5.

Demographics. The students also filled out demographic questions, asking for their gender, race, class-year, major, and GPA as well as perceptions of the computer image (how natural the video looked on a 5-point scale, where *1* = "extremely unnatural" and *5* = "extremely natural," as well as how attractive the professor was, rated from *1*= "extremely unattractive" to *5* = "extremely attractive.")

Procedure

The between-participants design, a 2 (race of professor: African American or White) X 2 (gender of professor) X 2 (gender of the student), had two independent variables (race and gender of the professor) and one quasi-independent variable (gender of the student).

After completing an informed consent form, students were assigned to condition using block randomization within student gender. Block randomization was used to ensure that relatively equal numbers of women and men would view each of the four "professors." Participants were told that the study was created to examine the effectiveness of online-teaching simulations. Each participant was shown the lecture on a computer in an individual cubicle while wearing headphones. After watching the lecture, the student filled out an online teacher evaluation form, followed by the 10-item true/false quiz. Finally, students were asked questions about the computer image (its naturalness and attractiveness) as well as personal demographics. At the conclusion of the study, the participants were asked what they thought the study was about before they were debriefed. No one guessed the race and gender variables.

Results

Preliminary Analyses

In examining participants' ratings of how natural the computer animation appeared, no significant main effect of the independent variables was found, although the mean ratings indicated that students perceived the computerized image to be very unnatural ($M = 1.40$, $SD = .70$. Because there was a significant main effect for professor gender on

Table 1. Correlations of Evaluations and Quiz Scores, GPA, Attractiveness, and Naturalness Ratings ($N = 329$)

Factor	Quiz Score	GPA	Attractiveness	Naturalness
Scholarship	.08	-.10	.16**	.17**
Organization/Clarity	.14*	-.09	.21**	.19**
Instructor-Group	.09	-.00	.27**	.28**
Instructor-Individual	.08	-.04	.29**	.28**
Dynamism/Enthusiasm	.09	.00	.25**	.26**
Overall	.13*	-.07	.22**	.29**
Quiz Score		.11	-.05	-.03

Note. *p < .05, **p < .01

attractiveness ratings of the computer-animated professor, as noted previously, and because both attractiveness and naturalness ratings were significantly positively correlated with all the student ratings measures (see Table 1), these two ratings (attractiveness and naturalness) were used as covariates in all analyses.

An analysis of student GPA revealed that female students had significantly higher GPAs ($M = 3.38$, $SD = .36$) than male students ($M = 3.22$, $SD = .41$) $F(1, 295)=13.30$, $p < .001$, $\eta 2 = .043$, especially in the woman professor conditions, $F(1, 295) = 6.28$, $p = .013$, $\eta 2 = .031$. Therefore, GPA also was used as a covariate in all analyses.

Student Ratings

The six dependent variables measured by the teacher evaluation form (five factor scores plus the overall rating) were examined in a 2 X 2 X 2 MANCOVA.

Contrary to Hypothesis 1, there was no significant interaction between professor gender and student gender. However, there was a significant MANCOVA main effect of student gender, $F(6, 284) = 3.39$, $p = .003$, $\eta^2 = .067$, observed power = .939, with male students rating professors higher than did female students on the following measures: Instructor-Group Interaction, Instructor-Individual Student Interaction, Dynamism Enthusiasm, and overall. Results approached significance on Organization/Clarity. See Table 2 for F values, effect sizes, means and SDs. There was no significant effect of student gender on the Scholarship factor.

Professor race had a significant MANCOVA effect on student ratings, $F(6, 284) = 2.61$, $p = .018$, $\eta^2 = .052$, but in a direction opposite that predicted in Hypothesis 2. African American professors were rated significantly higher than White professors on Instructor-Group interaction, $F(1, 289) = 5.05$, $p = .025$, $\eta^2 = .017$, observed power = .61. Findings approached significance on ratings of Instructor-Individual Student Interaction, $F(1, 289) = 3.18$, $p = .076$, $\eta 2 = .011$, observed power = .428. See Table 3 for means and SDs.

Quiz Scores

A 3-way ANCOVA on quiz scores revealed main effects of professor gender, $F(1, 290) = 4.89$, $p = .028$, $\eta^2 = .017$, observed power = .596, and professor race, $F(1, 290) = 5.98$, $p = .015$, $\eta^2 = .02$, observed power = .683. Participants who viewed male professors scored significantly higher on the quiz ($M = 6.64$, $SD = 1.46$) than did those who viewed female professors ($M = 6.18$, $SD = 1.56$). Similarly, participants who viewed White professors scored significantly higher on the quiz ($M = 6.64$, $SD = 1.46$) than did those who viewed

Table 2. Means, Standard Deviations (SD), and F Values of Ratings of Professors and Quiz Scores by Student Gender

Measure	Student Gender				$F(1, 289)$	η^2
	Men ($N = 116$)		Women ($N = 184$)			
	M	*SD*	*M*	*SD*		
Scholarship	3.08	.71	3.01	.69	.68	.00
Organization/Clarity†	3.66	.70	3.49	.77	3.63	.012
Instructor-Group**	2.68	.77	2.50	.86	7.44	.025
Instructor-Individual*	2.62	.77	2.54	.99	4.14	.014
Dynamism/Enthusiasm**	3.04	.86	2.81	.88	9.48	.032
Overall**	2.91	.91	2.56	1.06	13.23	.044
Quiz Scores	6.48	1.34	6.36	1.63		
Attractiveness	2.52	1.03	2.70	.98		
Naturalness	1.36	.70	1.43	.70		

Note: † $p < .10$, * $p < .05$, ** $p < .01$

African American professors (M = 6.17, *SD* = 1.56).

There was no main effect for student gender nor any significant interactions. Student performance did not significantly differ by student major (Humanities, Social Sciences, Natural Sciences, Engineering, or Interdisciplinary).

As shown in Table 1, scores on the quiz were significantly correlated ($p < .05$) only with student ratings of Organization/Clarity, $r(326) = .14$, and overall effectiveness, $r(326) = .13$. Students who rated the computer-animated professor higher in overall effectiveness as well as on questions tapping organization and clarity were more likely to score higher on the quiz than students who rated the professor low on these qualities.

Discussion

Overall, although the hypotheses were not supported when student ratings of teaching effectiveness were examined, the hypotheses were supported when student performance was considered. Specifically, student evaluations did not vary by teacher gender, nor was there a significant interaction with student gender (Hypothesis 1), but students who viewed a male professor scored better on the quiz than did students who viewed a female professor (Hypothesis 3). Similarly, although students rated the African American professor higher than the White professor on several teaching dimensions, those tapping hypothetical interactions with students, directly opposite prediction (Hypothesis 2), students who had an African American professor actually performed more poorly on the quiz than did students who had a White professor (Hpothesis 3). Another unexpected result was that student evaluation ratings by male students were significantly higher than female students. As predicted, student quiz performance was only slightly (but significantly) related to overall ratings of teacher effectiveness and to ratings on Organization/clarity.

Male students were expected to rate male professors higher than female professors while female students were expected to show the opposite pattern (Hypothesis 1), based on previous research (Bachen et al., 1999; Basow, 1998.) Such a pattern was not found, perhaps because male students tended to give all professors higher ratings than did female

Table 3. Means and Standard Deviations of Ratings of Professors and Quiz Scores by Professor Race

Measure	Professor Race			
	White ($N = 152$)		Black ($N = 148$)	
	M	*SD*	*M*	*SD*
Scholarship	2.98	.72	3.09	.66
Organization/Clarity	3.59	.71	3.52	.78
Instructor/Group*	2.48	.78	2.66	.86
Instructor/Individual†	2.42	.84	2.59	.90
Dynamism/Enthusiasm	2.87	.87	2.94	.89
Overall	2.66	.99	2.74	1.05
Quiz Scores*	6.64	1.46	6.17	1.56
Attractiveness	2.64	.98	2.63	1.02
Naturalness	1.42	.70	1.38	.70

Note: † $p < .10$. * $p < .05$

students on all teaching factors except Scholarship. (The lack of significance on Scholarship may be due to its relatively low internal reliability–.68–in this study.) The absence of the predicted gender interaction may be a sign that gender bias is less a factor today than it was previously, given the greater percentage of women in the professorate today (42%; National Center for Educational Statistics, 2007.) The unexpected main effect for student gender is inconsistent with previous research (e.g., Tatro, 1995). Perhaps the technical topic appealed more to men more than to women. Student major did vary by gender (more male than female engineering and business/economics majors, more female than male psychology and biology majors), but students with different majors did not significantly vary in how they rated professors. Another possible explanation for the higher ratings given by male participants may be the mode of administration: men may like information presented on a computer or by an avatar more than do women (Luppicini, 2007). These results, if replicated, may have implications for on-line instruction in general.

The higher ratings received by African American lecturers compared to White lecturers are opposite those hypothesized (Hypothesis 2) as well as the results of previous research, both naturalistic and laboratory (Bavishi et al., 2010; Ho et al., 2009; Reid, 2010; Smith, 2007). However, the only significant finding was with respect to how the professor would likely interact with students as a group. Because this rating was hypothetical, it may have been more vulnerable than other ratings (such as of organization or dynamism) to subjective factors. For example, it is possible that students over-valued the African American professors in order to appear non-prejudiced, as Harber (1998) found when White students had to give feedback to a peer they thought was African American. In Harber's study, the differential feedback only occurred with respect to subjective factors, such as essay quality, and not more objective ones, like essay mechanics.

Unlike the student ratings of teacher effectiveness, the quiz results were more consistent with predictions (Hypothesis 3). Those who saw a White professor and those who saw a

male professor scored higher on the quiz than those who saw an African-American professor and those who saw a female professor. The gender results are consistent with those found by Basow (1990). It may be that students pay less attention to, or put less credence on, the lecture of a non-normative professor than to the more normative one, especially since the lecture topic was a technical ("masculine") one. Because performance on the quiz is less subject to socially desirable responding than are student ratings of teaching, quiz performance may be the more sensitive measure of student reactions to professor race and gender.

The current results add to literature questioning the validity of student ratings of professors (Addison et al., 2006; Clayson, 2009; Greenwald & Gillmore, 1997; Zabaleta, 2007). There were very few significant correlations between quiz performance and student ratings of teaching effectiveness, and the two that did occur (for Organization/clarity and overall effectiveness) were very small in size. Furthermore, attractiveness ratings significantly correlated with each of the teaching dimensions, as other researchers have noted (e.g., Hamermesh & Parker, 2005; Riniolo, Johnson, Sherman, & Misso, 2006), which further undermines our confidence that student ratings actually assess teaching prowess. Nonetheless, the two significant correlations do suggest that those who learned more view their instructor as better organized and more effective than those who learned less.

One of the limitations of the study was the use of a very brief (three minute) computer-animated lecture given by a talking head that was considered to be very unnatural-looking. Clearly, classroom professors present a very different stimulus: they use a variety of teaching styles and visual aids, and they interact with students both in and outside of class. Nonetheless, significant results were found despite using naturalness ratings (as well as attractiveness ratings and GPA) as covariates. Furthermore, Ambady and Rosenthal (1993) have shown that students form a general impression of an instructor very quickly, based on rather limited information; this general impression then forms the basis of ratings of teacher effectiveness.

Another limitation of the study was the predominantly White sample. Because minority students were relatively few in number, separate analyses by race/ethnicity were not possible. Although results did not change much when the results from only White students were analyzed, there may be interactions between professor race and student race (Dee, 2005; Ehrenberg, Goldhaber, & Brewer, 1995; Gramzow & Gaertner, 2005) that should be considered in future research. Clearly, these results need to be replicated with diverse populations using a variety of lecture topics.

Finally, although laboratory studies have benefits over naturalistic studies given the greater degree of control over potential confounds, the effect sizes of all findings are small, which suggests that their practical influence may be difficult to discern in the field. The effects of professor race and gender on student ratings and achievement are likely to be subtle, affected by many other variables (student race and gender, course content and discipline, personality traits, grading leniency, etc.)

Overall, the current study suggests that student evaluations may not be a good indicator of actual teaching effectiveness. Students may rate professors highly even when they do not seem to learn from them, as suggested by the higher student evaluations received by African American professors but the lower quiz performance of their students. Gender bias in student ratings may be harder to discern than in the past, but may still operate at the level of student attention and interest and show up in student performance. Clearly, student evaluations need to be used cautiously, with recognition of their many limitations.

References

Addison, W. E., Best, J., & Warrington, J. D. (2006). Students' perceptions of course difficulty and their ratings of the instructor. *College Student Journal, 40,* 409-416.

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,* 431-441.

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. (1990). Effects of teacher expressiveness: Mediated by teacher sex-typing? *Journal of Educational Psychology, 82,* 599-602.

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

Basow, S. A. (1998). Student evaluations: The role of gender bias and teaching styles. In L.H. Collins, J.C. Chrisler, & K. Quina (Eds.), *Career strategies for women in academe: Arming Athena* (pp. 135-156). Thousand Oaks, CA: Sage.

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

Basow, S. A., & Distenfeld, M. S. (1985). Teacher expressiveness: More important for male teachers than female teachers? *Journal of Educational Psychology, 77,* 45-52.

Basow, S. A., & Martin, J. L. (2012). Bias in student ratings. In M.E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 40-49). Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/evals2012/index.php

Basow, S. A., & Montgomery, S. (2005). Student evaluations of professors and professor self-ratings: Gender and divisional patterns. *Journal of Personnel Evaluation in Education, 18,* 91-106.

Basow, S. A., Phelan, J., & 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.

Bavishi, A., Madera, J. M., & Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education, 3,* 245-256.

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluations. *Journal of Educational Psychology, 74,* 170-179.

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58,* 1019-1027.

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,* 855-868.

Biernat, M., & Kobrynowicz, D. (1997). Gender- and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality & Social Psychology, 72,* 544-557.

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 25-44). Bolton, MA: Anker.

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

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

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review, 95,* 158-165.

DuCette, J., & Kenney, J. (1982). Do grading standards affect student evaluations of teaching? Some new evidence on an old question. *Journal of Educational Psychology, 74,* 308-314.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Bulletin, 108,* 233-256.

Ehrenberg, R. G., Goldhaber, D.D., & Brewer, D.J. (1995). Do teachers' race, gender, and ethnicity matter? Evidence from the national educational longitudinal study of 1988. *Industrial and Labor Relations Review, 48,* 547-561.

Feldman, K. (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,* 151-211.

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

Gramzow, R. H., & Gaertner, L. (2005). Self-esteem and favoritism toward novel in-groups: The self as an evaluative base. *Journal of Personality and Social Behavior, 88,* 801-815.

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

Hamermesh, D. S., & Parker, A. M. (2005). Beauty in the classroom: Professors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24,* 369-376.

Harber, K. D. (1998). Feedback to minorities: Evidence of a positive bias. *Journal of Personality and Social Psychology, 74,* 622-628.

Hildebrand, M. & Wilson, R. C. (1970). *Effective university teaching and its evaluation.* Berkeley, CA: University of California Center for Research and Development in Higher Education.

Ho, A. K., Thomsen, L., & Sidanius, J. (2009). Perceived academic competence and overall job evaluations: Students' evaluations of African American and European American professors. *Journal of Applied Social Psychology, 39,* 389–406.

Leventhal, L., Perry, R. P., & Abrami, P. C. (1977). Effects of lecturer quality and student perception of lecturer's experience on teacher ratings and student achievement. *Journal of Educational Psychology, 69,* 360-374.

Luppicini, R. (2007). Review of computer mediated communication research for education. *Instructional Science, 35,* 141–185.

Millea, M., & Grimes, P. W. (2002). Grade expectations and student evaluation of teaching. *College Student Journal, 36,* 582 - 590.

National Center for Educational Statistics. (2007). [Table showing demographics of full-time instructional faculty by race/ethnicity, sex, and academic rank]. *Digest Of Education Statistics,* Table 249. Retrieved from http://nces.ed.gov/programs/digest/d08/tables/dt08_249.asp

Phelan, J. E., Moss-Racusin, C. A., & Rudman, L. A. (2008). Competent yet out in the cold: Shifting criteria for hiring reflect backlash toward agentic women. *Psychology of Women Quarterly, 32,* 406-413.

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education, 3,* 137-152.

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *The Journal of General Psychology, 133,* 19-35.

Seldin, P. (1999). Current practices-good and bad nationally. In P. Seldin (Ed.). *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 1-24). Boston, MA: Anker.

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin, 26,* 1329-1342.

Smith, B. P. (2007). Student ratings of teaching effectiveness: An analysis of end-of-course faculty evaluations. *College Student Journal, 41,* 788-800.

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

Tatro, C. N. (1995). Gender effects on student evaluations of faculty. *Journal of Research & Development in Education, 28,* 169-173.

Timmerman, T. (2008). On the validity of RateMyProfessors.com. *Journal of Education for Business, 84,* 55– 61.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education, 12,* 55-76.