The Journal of Higher Education

ISSN: 0022-1546 (Print) 1538-4640 (Online) Journal homepage: www.tandfonline.com/journals/uhej20

Is There Gender Bias in Student Evaluations of Teaching?

John A. Centra & Noreen B. Gaubatz

To cite this article: John A. Centra & Noreen B. Gaubatz (2000) Is There Gender Bias in Student Evaluations of Teaching?, The Journal of Higher Education, 71:1, 17-33, DOI: 10.1080/00221546.2000.11780814

To link to this article: https://doi.org/10.1080/00221546.2000.11780814

John A. Centra
Noreen B. Gaubatz

Is There Gender Bias in Student Evaluations of Teaching?

Given the widespread use of student evaluations of teaching for tenure and promotion decisions, it is important to be aware of possible bias in the evaluations. One definition of bias is if a teacher or course characteristic affects teacher evaluations, either positively or negatively, but is unrelated to criteria of good teaching, such as increased student learning. Class size, for example, affects evaluations in that teachers of classes with under 15 students get higher evaluations. But if students learn more in small classes than they do in large classes, perhaps because small classes allow for more personal attention, then class size is not truly biasing the evaluations; rather, the evaluations are reflecting increased learning.

A second, more general, definition of bias is when a known characteristic of students systematically affects their ratings of teachers. The gender of the student, particularly how it interacts with the gender of the teacher, is an example of this possible bias in student evaluations. Do male students tend to rate women teachers lower than men teachers because of a gender bias, especially in fields that are male dominated, such as the natural sciences? Do female students judge women teachers to be more effective than men because they feel more comfortable with them? These are important questions that directly affect the validity of the evaluations when used for personnel decisions. Ideally student evaluations

The authors wish to acknowledge the support of the Higher Education Assessment Program at the Educational Testing Service for providing the data for this study. The results were also published as a Student Instructional Report II publication.

John A. Centra is research professor and professor emeritus, and Noreen B. Gaubatz is a doctoral student in higher education at Syracuse University.

The Journal of Higher Education, Vol. 70, No. 1 (January/February 2000)

should be related to what they learn from a teacher and not to gender or to other personal characteristics of the teacher (e.g., age, ethnicity).

Studies that have investigated gender bias have thus far produced conflicting results. Some studies have found no (or extremely small) differences between the evaluation of female and male instructors on the basis of student gender alone (Basow & Distenfeld, 1985; Basow & Howe, 1987; Bennett, 1982; Elmore & LaPointe, 1974; Harris, 1975; Kaschak, 1981). Other studies reported gender bias, with male students rating female instructors lower than male instructors (Basow & Silberg, 1987; Etaugh & Riley, 1983; Kaschak, 1978; Lombardo & Tocci, 1979; Paludi & Bauer, 1983).

Two studies conducted in actual classrooms did not report gender bias in overall evaluations. Bennett's (1982) used a course evaluation questionnaire that included teaching performance ratings, perceptual orientation scales, and indicators of the degree and context of student-instructor interaction. Her data included the evaluations of 11 female and 28 male instructors by 253 students enrolled in nonscience introductory courses at a liberal arts college. Female and male students did not differentiate between faculty members of different gender. Although there was no evidence of direct bias in formal student evaluations of instructors, there was evidence of gender related differences in regard to student-instructor relationships and instructor warmth, support, and accessibility. Elmore and LaPointe (1974) found no interaction between faculty gender and student gender in their analysis of 38 pairs of courses (paired on the basis of course number and gender of instructor) evaluated by 1,259 students. The courses were selected from a variety of departments and colleges within one research university. No attempts were made, however, to compare results from various disciplines or results for students in the same classes.

Research that reported gender bias included Basow and Silberg's (1987) study of 16 pairs of instructors (paired on rank, discipline, and years of experience), in which male students rated female instructors less favorably than male instructors. Similar results by Kaschak (1978) and Lombardo and Tocci (1979) were found in a simulated rather than actual classroom setting. In these studies female students saw no difference in the effectiveness of male and female teachers.

Feldman conducted two reviews of students' views of male and female college teachers. In the first review, results of laboratory studies (the use of photographs, descriptions, simulations, etc.) indicated that "little, if any, same- gender or cross-gender bias is evidenced" (Feldman, 1992, p. 359). The second review summarized results from studies of actual classrooms. Ten studies examined an overall rating of teachers for which there was a slight tendency toward same- gender preference

(Feldman, 1993, p. 169). But as Feldman noted, the difference among means was not always statistically significant. Male and female students were in different classes in some of the studies, and other variables such as the academic area of the course were not examined (or examined in only two studies where only a few fields were included and the results conflicted). Thus, combining individual studies with small *n*'s and uncontrolled variables did not produce conclusive results.

Clearly the question of gender bias in the evaluations of instructors has not been fully resolved. The study reported here addressed many of the limitations found in the current research. Using actual student ratings of classroom instructors instead of a simulated design increased the utility of the results. A large number of different types of institutions (two-year and four-year colleges and universities) were included, whereas many of the past studies were limited to a single institution (Basow & Silberg, 1987; Bennett, 1982; Elmore & LaPointe, 1974). A variety of academic disciplines, rather than one or a relatively small number of disciplines, were analyzed. And finally, unlike in previous studies, the class rather than the individual student was the unit of analysis. Mean ratings of male and female students in each of a large number of classes were analyzed, thereby increasing reliability and minimizing the effects of extraneous student and teacher variables other than gender.

Method

This study examined gender differences through two analyses. In the first female and male student ratings in the same classes were compared for female instructors and for male instructors. This analysis more directly addressed the purpose of this study because it compared ratings by students of the same instructors. Figure 1 illustrates this analysis.

In the second analysis the ratings by all male students were examined for how they differed for male and female instructors. The ratings by all female students were compared in the same way. Figure 2 illustrates this

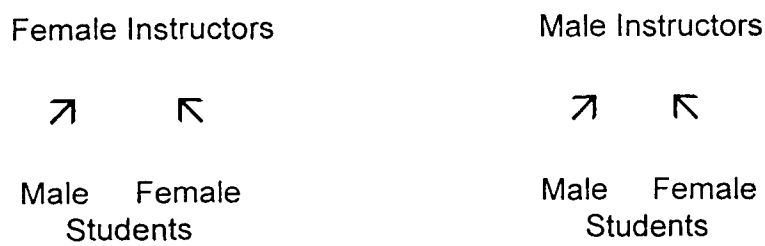

FIG. 1. Student Ratings Within Classes

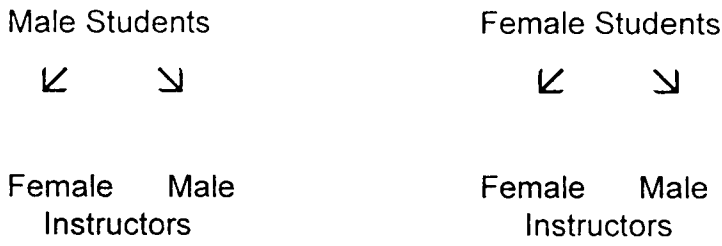

FIG. 2. Student Ratings Across Classes

analysis. Because the male or female students have evaluated different instructors, this analysis focuses more on student gender effects and their interactions with instructor gender effects across classes.

Subjects

The data in this study included 741 classes, each of which had an enrollment of at least 10 female students and 10 male students. The minimum number of students was based on Centra's (1998) research, indicating that scores based on 10 or more students provide a sufficient level of reliability for research purposes. Specifically, the intraclass reliabilities were in the high 50s for *individual items*; the *scale means* used in this study would have higher intraclass reliabilities for ten students. Moreover, the coefficient alpha and test-retest reliabilities for the scale scores were generally above 0.90 and 0.80, respectively. Student evaluation forms were administered over three semesters—spring and fall semesters of 1995, and spring semester of 1996. The sample of 21 institutions included similar numbers of two- and four-year colleges and universities, although about half of the classes were from universities (largely comprehensive rather than research or doctoral universities). Table 1 indicates the breakdown by institutional type for this study. For analysis purposes, the classes were collapsed into the following eight discipline groups: health sciences (20 classes), business (75 classes), education (43 classes), social sciences (161 classes), fine arts (45 classes), natural sciences (167 classes), technology (32 classes), and humanities (173 classes). Female instructors taught 211 (28%) of the classes and male instructors taught 530 (72%) of the classes. Only in the health sciences were there more female than male instructors.

Instrument

The student evaluation form used in this study was the Student Instructional Report II (SIR II), a new version of the original SIR, which

TABLE 1
Institutional Type

Institutional Type	Number of Type	Percentage of Classes from Each Type
2-year college	7	26%
4-year college	8	25%
University	6	49%

had been made available to colleges over the past 25 years by the Educational Testing Service (ETS). SIR II was developed to reflect more recent emphases in college teaching, such as active learning, course outcomes assessment, and the importance of student effort and involvement (Centra, 1998). This end-of-course survey consists of seven scales: Course Organization and Planning (5 questions); Communication (5 questions); Faculty/Student Interaction (5 questions); Assignments, Exams, and Grading (6 questions); Course Outcomes (5 questions); Student Effort and Involvement (3 questions); and Course Difficulty, Workload, and Pace (3 questions).

Course Organization and Planning included such items as the instructor's explanation of course requirements, use of class time, and way of summarizing or emphasizing important points in class. Communication included the instructor's ability to make clear and understandable presentations, use of challenging questions or problems, and enthusiasm for course material. Within the Faculty/Student Interaction scale were such items as the instructor's helpfulness and responsiveness to students, concern for student progress, availability for extra help, and willingness to listen to student questions and opinions. Items in the Assignments, Exams, and Grading scale included the information given to students about how they would be graded, the clarity of exam questions, the instructor's comments on assignments and exams, and the helpfulness of assignments in understanding course materials. Students responded to these items as practices that ranged from "Very Effective" to "Ineffective" (5-point scale) in contributing to their learning in the course. A 5-point scale was also used for the Course Outcomes and Student Effort and Involvement scales, with 5, the top response, being "much more than most courses," and 1, the bottom response, indicating "much less than most courses." Course Outcomes, the only scale that assessed student perceptions of learning, included the students' ratings of progress toward course objectives, increase in learning and interest in the subject matter, whether the course helped students to think independently about

the subject matter, and whether the course actively involved them in what they were learning. In the Student Effort and Involvement scale students reported the extent to which they put effort into the course, were prepared, and were challenged. The final scale included three items that rated the Course Difficulty, Workload, and Pace, with the top response of three being "about right." For the first six scales, a pilot test indicated that the response options described above had better statistics and were more useful than an agree/disagree response format.

The development of these scales is described in the manual (Centra, 1998). An overall evaluation item, seven items that rated supplementary instructional methods (not a scale), and a student information section (five questions) are also included in the 45-item form.

Instructors completed a "Cover Sheet" that included two questions that were used in this analysis: how the class was conducted (e.g., lecture, discussion, etc.), and the class size (in categories such as 16–35).

Data Analysis

Multivariate analysis of variance (MANOVA) was used to investigate the mean differences on the several dependent variables simultaneously while controlling for the intercorrelation among them. By considering all the variables simultaneously, MANOVA becomes more powerful than performing separate ANOVAs on each of the variables. Feldman (1992) pointed out that in general the studies he reviewed on this topic did not control for multiple *F*-test errors. A measure of effect size was also computed to address the practical utility of the findings.

Results

MANOVA results for the class mean scores on the seven SIR II scales and the overall evaluation item are presented in Tables 2 and 3. Separate MANOVAs were generated for female instructors, male instructors, female students, male students, and the interaction between instructor gender and student gender across each discipline and for all disciplines combined. Follow-up ANOVA tests were examined to determine on which scales instructors were rated differently based on student and instructor gender. Table 4 contains the mean ratings of female and male instructors by female and male students. Asterisks identify significant ANOVAs for scales that indicate between instructor (gender) differences. Plus signs identify significant ANOVAs for scales that indicate between student (gender) differences. Results are noted at the 0.05 level of significance or better.

TABLE 2

MANOVA Results and Sample Sizes—Ratings of Female and Male College Instructors by Female and Male Students Within the Same Classes

	Sample Size		Instructor Gender	
	Female Instructor	Male Instructor	Female Instructor	Male Instructor
All disciplines	$N = 211$	$N = 530$	2.61**	6.88***
Health	$N = 14$	$N = 6$	0.27	2.45
Business	$N = 20$	$N = 55$	2.13	1.42
Education	$N = 9$	$N = 34$	0.89	1.46
Social sciences	$N = 38$	$N = 123$	1.45	2.67**
Fine arts	$N = 16$	$N = 29$	1.44	1.68
Natural sciences	$N = 40$	$N = 127$	1.84	2.09*
Technology	$N = 8$	$N = 24$	0.69	1.15
Humanities	$N = 57$	$N = 116$	1.13	1.85

NOTE: MANOVA results are represented by Wilks' Lambda.
*$p < = 0.05$. **$p < = 0.01$. ***$p < = 0.001$.

TABLE 3

MANOVA Results and Sample Sizes—Ratings by Female and Male Students of Instructors Across Different Classes

		Student Gender		Instructor Gender x
	Sample Size	Female Student	Male Student	Student Gender
All disciplines	$N = 741$	4.23***	5.62**	0.80
Health	$N = 20$	1.09	1.24	0.23
Business	$N = 75$	2.17*	0.54	0.51
Education	$N = 43$	0.63	1.38	0.22
Social sciences	$N = 161$	1.85	2.60*	1.13
Fine arts	$N = 45$	0.95	1.82	1.50
Natural sciences	$N = 167$	3.84***	3.91***	0.56
Technology	$N = 32$	0.88	1.07	0.30
Humanities	$N = 173$	1.45	2.12*	0.64

NOTE: MANOVA results are represented by Wilks' Lambda.
*$p < = 0.05$. **$p < = 0.01$. ***$p < = 0.001$.

Analysis Within the Same Classes

The first analysis examined the question, Do female and male students in the same classes rate their instructors differently depending on the gender of the instructor? MANOVA results are obtained from Table 2 and ANOVA results/mean scores are obtained by reading the vertical columns of Table 4 (plus signs indicate statistical significance). The mean scores being compared in this question are from female and male students within the same classes; thus the two groups of students are evaluating the same instructors.

TABLE 4

Mean Ratings of Male and Female Instructors by Male and Female College Students (ANOVA Results)

Discipline		Scale A Course Organization & Planning MI		Scale A FI		Scale B Communication MI		Scale B FI		Scale C Faculty/Student Interaction MI		Scale C FI		Scale D Assignments, Exams, & Grading MI		Scale D FI		Scale F Course Outcomes MI		Scale F FI		Scale G Student Effort & Involvement MI		Scale G FI		Scale H Course Diff., Workload & Pace MI		Scale H FI		Overall Evaluation MI		Overall Evaluation FI	
All disciplines	MS	4.24	*	4.16	++	4.27		4.21	++	4.19		4.21	+	3.96		4.00	++	3.64		3.61		3.54		3.50	+	2.54		2.53		3.95		3.92	+
	FS	4.26		4.28		4.29		4.32		4.21	**	4.31		3.98	***	4.10		3.58		3.65		3.59		3.59		2.53		2.55		4.00		4.04	
Health	MS	4.09		3.93		4.32		4.08		4.34		3.91		3.99		3.86		3.92		3.64		3.55		3.67		2.63		2.27		3.85		3.70	
	FS	4.03		4.07		4.35		4.22		4.29		3.93		4.00		3.93		4.00		3.76		3.57		3.76		2.56		2.34		3.91		3.89	
Business	MS	4.19		4.25		4.23		4.27		4.15		4.27		3.94		4.09		3.65		3.77		3.58		3.55		2.57		2.61		3.92		4.00	
	FS	4.12		4.36		4.18		4.38		4.07		4.36		3.86	**	4.18		3.47	*	3.71		3.57		3.49		2.50		2.54		3.84		4.03	
Education	MS	4.39		4.15		4.41		4.27		4.47		4.42		4.19		4.07		3.89		3.68		3.47		3.56		2.58		2.60		4.09		3.78	
	FS	4.28		4.17		4.37		4.27		4.43		4.40		4.11		4.12		3.81		3.69		3.43		3.51		2.59		2.64		4.12		3.94	
Social science	MS	4.34	**	4.11		4.33	*	4.17		4.22		4.15		4.00		3.91		3.66		3.52		3.52		3.40		2.61		2.59		4.02		3.88	
	FS	4.34		4.22		4.33		4.27		4.25		4.26		4.01		3.97		3.61		3.55		3.58		3.51		2.60		2.60		4.06		3.90	
Fine arts	MS	4.15		3.97		4.22	*	4.03	+	4.07		4.03	+	3.97		3.86		3.54		3.39		3.32		3.14		2.48		2.55		3.78	+	3.72	
	FS	4.32		4.26		4.37		4.30		4.31		4.22		4.06		4.08		3.60		3.50		3.40		3.30		2.54		2.65		4.03		3.92	
	MS	4.17		4.20		4.19		4.22		4.10	*	4.28		3.85	*	4.02		3.49		3.57		3.65		3.60		2.39		2.40		3.88		4.02	

TABLE 4 (*Continued*)

Discipline		Scale A Course Organization & Planning MI	FI	Scale B Communication MI	FI	Scale C Faculty/Student Interaction MI		FI	Scale D Assignments, Exams,& Grading MI		FI	Scale F Course Outcomes MI		FI	Scale G Student Effort & Involvement MI	FI	Scale H Course Diff., Workload & Pace MI	FI	Overall Evaluation MI		FI
	MS	4.17	4.20	4.19	4.22	4.10	*	4.28	3.85	*	4.02	3.49		3.57	3.65	3.60 +	2.39	2.40	3.88		4.02
Natural science																					
	FS	4.21	4.32	4.22	4.33	4.12	***	4.40	3.88	***	4.17	3.44	*	3.64	3.73	3.82	2.39	2.41	3.91	*	4.13
	MS	3.89	4.02	3.98	4.04	3.94		4.10	3.75		3.99	3.56		3.72	3.56	3.32	2.41	2.43	3.69		3.84
Technology																					
	FS	3.93	4.19	4.04	4.13	4.04		4.18	3.85	*	4.19	3.56		3.73	3.66	3.64	2.41	2.42	3.85		4.05
	MS	4.27	4.21	4.32	4.26 +	4.24		4.26	4.00		4.06	3.70		3.66	3.49	3.54	2.62	2.62	4.05		3.95 +
Humanities																					
	FS	4.32	4.31	4.35	4.37	4.26		4.39	4.06		4.16	3.66		3.72	3.52	3.57	2.65	2.65	4.08		4.13

NOTE: Scale H was collapsed into a scale of 1–3. MI = male instructors; FI = female instructors; MS = male students; FS = female students.
*indicate significant ANOVAs for instructor gender differences.
+indicate significant ANOVAs for student gender differences.
$^{*}p \leq 0.05$. $^{**}p \leq 0.01$. $^{***}p \leq 0.001$.
$^{+}p \leq 0.05$. $^{++}p \leq 0.01$. $^{+++}p \leq 0.001$.

The two MANOVAs for all disciplines combined indicated that there are significant differences in how female and male instructors were evaluated by female and male students. The MANOVA for female instructors (F = 2.61) was significant at the 0.01 level, whereas the MANOVA for the male instructors (F = 6.88) was significant at the 0.001 level (Table 2). The ANOVA results indicated that when female instructors were rated by female and male students, there was a significant difference in the mean ratings on five of the seven scales and the overall evaluation item, with the female students consistently awarding higher scores to female instructors (Table 4). Mean scores for female instructors were: Course Organization and Planning (FS M = 4.28; MS M = 4.16); Communication (FS M = 4.32; MS M = 4.21); Faculty/Student Interaction (FS M = 4.31; MS M = 4.21); Assignments, Exams, and Grading (FS M = 4.10; MS M = 4.00); Student Effort and Involvement (FS M = 3.59; MS M = 3.50); and overall rating (FS M = 4.04; MS M = 3.92). This indicates, of course, that female instructors received a lower rating from male students on these same scales. When male instructors were rated by female and male students, there were no significant ANOVAs, thereby indicating no significant difference in how the students of both genders evaluated the male instructors. In summary, female instructors received higher ratings on six of the eight variables when rated by female students. Male instructors were not rated significantly different by female or male students. One of the scales in which there was no difference in how instructors were evaluated by students of either gender was the Course Outcomes Scale, which assesses student perceptions of their learning in a course.

Do these results have any practical utility? Taking the differences between means and dividing by the standard deviation provides an effect size measure. The standard deviation for the five scales averaged 0.43, while it was 0.50 for the overall rating item. The mean differences on the five scales and the overall item ranged from 0.09 to 0.12 (average of 0.107), which indicates about a fourth of a standard deviation difference as an effect size.

The pattern of results at the various academic discipline levels was more varied. The MANOVA for male instructors in the Natural Sciences (F = 2.09) was significant, indicating that male instructors were rated significantly different by female and male students on the vector representing the seven scales and the overall evaluation item. There were, however, no significant ANOVAs. This finding is possible because the multivariate test considers the correlation among the variables and the joint differences on all the variables, while the univariate tests do not consider correlations among the variables and determine differences on

each variable separately. The MANOVA for female instructors (F = 1.84) was not significant, thus indicating no differences between the ratings by female and male students. Under these circumstances, the ANOVAs should be interpreted cautiously, because as the number of calculated F ratios increases, so does the likelihood of obtaining a significant F by chance (one scale was significant at the 0.05 level). In summary, in the Natural Sciences there was generally little difference in how female and male students evaluated female and male instructors.

In the Social Sciences, there was a significant MANOVA for the ratings of male instructors by female and male students (F = 2.67), but not for the female instructors (F = 1.45). There were, however, no significant ANOVAs for the evaluation of male instructors by female and male students.

For the remaining six disciplines there were no significant MANOVAs for female or male instructors when rated by female and male students, although Fine Arts and Humanities reported some significant ANOVAs. In summary, female and male instructors did not differ significantly in their ratings by students in these six academic disciplines.

Analysis Across Different Classes

In the second analysis, the question examined was, In general (i.e., across classes) do female and/or male students tend to give different ratings to female and male instructors? MANOVA results are given in Table 3, and ANOVA results/mean scores are obtained by reading the horizontal rows of Table 4 (asterisks indicate statistical significance). The mean scores compared in this question are from two groups of female students—one group evaluated female instructors and the other male instructors. Also, the mean scores for two groups of male students were compared—one group evaluated female instructors and the other male instructors. Actually, compared for this question were mean scores based on class means rather than individual student ratings. Previous studies have generally used individual student ratings across classes.

The two MANOVAs for all disciplines combined indicated that there were significant differences in how female and male students evaluated instructors. Both the female students' MANOVA (F = 4.23) and the male students' MANOVA (F = 5.62) were significant at the 0.001 level. The ANOVA results indicated that when female students evaluated female and male instructors, there was a significant difference in the mean ratings on two of the seven scales. In the two cases, the female students gave a higher rating to the female instructors than to the male instructors on Faculty/Student Interaction (FI M = 4.31; MI M = 4.21); and Assignments, Exams, and Grading (FI M = 4.10; MI M = 3.98). Male students

differed significantly in their evaluation of female and male instructors on only one scale, giving male instructors higher ratings on Course Organization and Planning (MI M = 4.24; FI M = 4.16).

In analyzing results from the various academic disciplines, the MANOVAs for the Natural Sciences were significant for both female (F = 3.84) and male students (F = 3.91). Female students rated the group of female instructors significantly higher than the group of male instructors on three of the seven scales— Faculty/Student Interaction (FI M = 4.40; MI M = 4.12); Assignments, Exams, and Grading (FI M = 4.17; MI M = 3.88); and Course Outcomes (FI M = 3.64; MI M = 3.44)—and the overall evaluation item (FI M = 4.13; MI M = 3.91). Male students rated female instructors significantly higher than male instructors on two of the scales—Faculty/Student Interaction (FI M = 4.28; MI M = 4.10); and Assignments, Exams, and Grading (FI M = 4.02; MI M = 3.85). In summary, in the Natural Sciences both female and male students rated female instructors higher than male instructors in certain areas of instruction. These results for female instructors are noteworthy in that Natural Science is a traditionally male dominated field. Additionally, in this study only 24% of the instructors in the Natural Sciences were women.

In the Social Sciences, the MANOVA for the evaluation of instructors of both genders by male students (F = 2.60) was significant, indicating an overall significant difference in how male students evaluated female instructors in comparison to male instructors. The ANOVAs indicated that male students rated the group of male instructors higher than the group of female instructors on two scales—Course Organization and Planning (MI M = 4.34; FI M = 4.11); and Communication (MI M = 4.33; FI M = 4.17). The MANOVA examining how female students evaluated female and male instructors was not significant (F = 1.85). In summary for the Social Sciences, there were few significant differences between student genders; however, on two scales male students rated male instructors higher than female instructors.

An analysis of the Business discipline indicated that only the MANOVA for female students evaluating the group of female and the group of male instructors was significant (F = 2.17). The ANOVAs indicated that female students rated female instructors significantly higher than male instructors on two scales— Assignments, Exams, and Grading (FI M = 4.18; MI M = 3.86); and Course Outcomes (FI M = 3.71; MI M = 3.47). In summary, for Business, there were small differences, but they were in favor of female students rating female instructors higher than male instructors on two scales.

Humanities reported one significant MANOVA, indicating that male students evaluated female instructors, as a group, significantly different

than the group of male instructors (F = 2.12). There were, however, no significant ANOVAs.

The remaining four disciplines—Health, Education, Fine Arts, and Technology—reported no significant MANOVAs for female students' ratings of the group of female and the group of male instructors, or male students' ratings of female and male instructors, although Fine Arts and Technology reported some significant ANOVAs. In these four disciplines, then, the ratings by female students of female and male instructors did not differ significantly. This is also true for the ratings by male students.

The lack of significant interactions in instructor/student gender for any of the disciplines individually or for all disciplines combined indicates small cross-gender effects. While the effect size for the significant means for all disciplines combined were about the same as in the first analysis (about one-fourth of a standard deviation), it was around a half standard deviation for the three disciplines (natural sciences, social sciences, and business).

Classroom Teaching Method

A final analysis looked at whether male and female instructors conducted their classes differently and thus may have received different ratings because of this. As summarized in Table 5, according to their self-reports male instructors were almost twice as likely to lecture as female instructors (22.4% vs. 12.4%). On the other hand, discussion as a pedagogy was used more by female than male instructors (5.6% vs. 3.3%). These statistically significant differences (Chi-Square = 18.508, p = 0.002) indicate marked contrasts to preferred approaches to teaching, although the majority of both groups (55%) used a combination of lecture

TABLE 5

How Class was Conducted as Reported by Male and Female Instructors

	Male Instructors N = 577		Female Instructors N = 233	
	Frequency	Percent	Frequency	Percent
Lecture	129	22.4	29	12.4
Lecture/discussion	315	54.6	128	54.9
Discussion	19	3.3	13	5.6
Lecture/laboratory	72	12.5	47	20.2
Laboratory	9	1.6	6	2.6
Other	18	3.1	10	4.3
No reply	15	2.6	—	—

Chi-Square (5,795) = 18.508. p = 0.002.

and discussion. Given the possibility that differences in class sizes may have enabled more women to conduct discussions and men to lecture, a second analysis was run according to self-reported class sizes. Instructors had indicated whether their classes were under 15 in size, 16–35, 36–100, or over 100. Because this study only analyzed classes with at least 10 male and 10 female students, there were no classes in the under 15 category. In the remaining categories, as Table 6 indicates, no significant differences in class sizes were evidenced for male and female instructors (Chi-Square = 5.05, $p = 0.08$).

Discussion

Past study results have been inconclusive or inconsistent, most likely because of shortcomings in their designs. In his review of ten studies, Feldman (1993) noted a slight tendency for same-gender preferences on an overall evaluation item, but he also noted that these studies failed to control for important variables such as the course and discipline. In the first analysis of this study, in which only mean student ratings within the *same classes* were compared (Figure 1), female instructors received higher ratings from female students on six of eight variables, whereas male instructors received equal ratings from both male and female students. One of the variables that was significant was the overall evaluation item, which is often emphasized in personnel judgments. Other differences indicate that female students, relative to male students in the same classes, saw female instructors as better organized, better communicators, more interactive, and providing higher quality exams, assignments, and feedback to students. On the other hand, for the Course Outcomes scale, there were no same- or cross-gender differences. Although this scale does not measure actual student learning or achievement, it does at least measure student perceptions of the amount and type of

TABLE 6
Class Size Reported by Male and Female Instructors

	Male Instructors $N = 577$		Female Instructors $N = 231$	
Number of Students[a]	Frequency	Percent	Frequency	Percent
16–35	350	60.7	157	67.3
36–100	225	39.0	72	30.9
Over 100	2	0.3	2	0.9
No reply	0	0	2	0.9

Chi-Square (2,808) = 5.05. $p = 0.08$.
[a]Classes with fewer than 20 students were excluded from the study.

learning they received in the course. Thus, considering the first definition of bias—that bias is when a characteristic such as gender affects evaluations systematically but does not affect learning—we would conclude that there is bias in favor of female instructors by female students.

Feldman (1993) argued that favorable ratings may not be bias but rather a reflection of better teaching. The lack of higher evaluations in student perceived learning would not support better teaching by female instructors. Instead, as further analyses indicated, female instructors tended to teach differently; they lectured less than males and used discussions more (with similar class sizes). It may in fact be these differences in teaching style that caused female teachers to get higher ratings on the scales that reflected communication, interaction, and feedback, although this was only by female students.

Belenky, Clinchy, Goldberger, and Tarule (1986) emphasized that female instructors, as well as female students, are more receptive to a teaching methodology that values connection over separation, understanding and acceptance over assessment, and collaboration over debate. "Connected classrooms" provide an environment for growth and an acceptance of uncertainty, because knowledge evolves over time and experience. Connected teachers, according to Belenky et al., emphasize group work and discussions, and see their role as that of a facilitator. This methodology is in contrast to a more traditional, lecturing approach to teaching that Freire (1971) described as the "banking" method, in which the teacher's role is to "fill" the students by making deposits of information.

Do the differences found within classes (Figure 1) have practical utility? The effect size of about a fourth of a standard deviation suggests a modest difference. The mean differences between instructor genders ranged from 0.09 to 0.12 on the significant five scales and the overall rating. Comparing these gaps to the national comparative data for two- and four-year institutions produced by ETS (1998) translates into about a 10 percentile difference for teachers in the middle and upper decile rating ranges. As the ETS guidelines suggest, differences of 10 percentile points or less are not critical; at least 20 percentile points are recommended as a significant gap (Educational Testing Service, 1998).

The within-class results for each of the eight academic disciplines varied somewhat, but generally the differences were not highly significant or consistent in any of the disciplines. Thus the discipline category of the course, or at least those studied here, were not critical in gender/rating relationships.

The second analysis (Figure 2) was similar to previous studies that compared male and female student ratings across classes, where the stu-

dents of each gender evaluated different instructors, thus introducing another source of variance. It differed from those previous studies, however, in that the only ratings analyzed were mean scores from classes with at least 10 male and 10 female students. For all disciplines combined, this analysis revealed some same gender preferences for both female and male students. The three scales where this occurred may reflect teaching style preferences: female students saw female teachers as more interactive, providing feedback on exams, and the like (Faculty/Student Interaction, and Assignments, Exams and Grading Scales); male students saw males as better organized and more systematic teachers (Course Organization and Planning Scale). As these results and further analysis indicated, interaction (cross-gender) effects were not significant. Moreover, as in the first analysis, the effect size was only about a fourth of a standard deviation (about 10 percentile points).

Larger differences (about one-half of a standard deviation) were reflected within some of the disciplines, but not always in the expected direction. In particular, in Natural Science, a male dominated field, both male and female students gave female instructors higher ratings in certain areas of instruction. Faculty/Student Interaction and Assignments, Exams and Grading were two of these areas, suggesting that female teachers in Natural Science were more approachable and helpful. Support for this came from looking at how classes were conducted by this sample of Natural Science instructors: as with the combined disciplines, men more often lectured and the women used more discussion (Chi-Square = 10.126, $p = 0.04$), even though class sizes were not significantly different.

In Social Science and Business the results paralleled the general findings: same gender preferences on a few scales.

Is there Gender Bias in Student Evaluations of Teaching? The results reflect some same gender preferences, particularly in female students rating female teachers. But the differences in ratings, though statistically significant, are not large and should not make much difference in personnel decisions. Moreover the higher evaluations received by female teachers from females, and in some instances from males as well (Natural Sciences in particular), could well be due to differences in teaching styles. Women in this study were more likely than men to use discussion rather than a lecture method, and as a group they appear to be a little more nurturing to students, as also reflected in certain scales in this study.

References

Basow, S. A., & Distenfeld, M. S. (1985). Teacher expressiveness: More important for males than females? *Journal of Educational Psychology, 77,* 45–52.

Basow, S. A., & Howe, K. G. (1987). Evaluations of college professors: Effects of professors' sex-type, and sex, and students' sex. *Psychological Reports, 60,* 671–678.

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.

Belenky, M. F., Clinchy, B. M., Goldberger, N. R., & Tarule, J. M. (1986). *Women's ways of knowing.* New York: Basic Books.

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.

Centra, J. A. (1998). *Development of The Student Instructional Report II.* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1998) *Student Instructional Report II, comparative Data, 1995–1997. For four-year colleges and universities, for two-year colleges,* Princeton, NJ, Higher Education Assessment Program.

Elmore, P. B., & LaPointe, K. A. (1974). Effects of teacher sex and student sex on the evaluation of college instructors. *Journal of Educational Psychology, 66,* 386–389.

Etaugh, C., & Riley, S. (1983). Evaluating competence of women and men: Effects of marital and parental status and occupational sex-typing. *Sex Roles, 9,* 943–952.

Feldman, K.A. (1992). College students' views of male and female college teachers: Part 1—Evidence from the social laboratory and experiments. *Research in Higher Education, 33,* 317–351.

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

Freire, P. (1971). *Pedagogy of the oppressed.* New York: Seaview.

Harris, M. B. (1975). Sex role stereotypes and teacher evaluations. *Journal of Educational Psychology, 67,* 751–756.

Kaschak, E. (1978). Sex bias in student evaluations of college professors. *Psychology of Women Quarterly, 2,* 235–243.

Kaschak, E. (1981). Another look at sex bias in students' evaluations of professors: Do winners get the recognition that they have been given? *Psychology of Women Quarterly, 5,* 767–772.

Lombardo, J., & Tocci, M. (1979). Attribution of positive and negative characteristics of instructors as a function of attractiveness and sex of instructor and sex of subject. *Perceptual and Motor Skills, 48,* 491–494.

Paludi, M. A., & Bauer, W. D. (1983). Goldberg revisited: What's in an author's name. *Sex Roles, 9,* 387–390.