J Pers Eval Educ (2005) 18: 91–106
DOI 10.1007/s11092-006-9001-8

Student Ratings and Professor Self-Ratings of College Teaching: Effects of Gender and Divisional Affiliation

Susan A. Basow · Suzanne Montgomery

Published online: 3 June 2006

Abstract Twenty female and 23 male professors at a liberal arts college participated along with their 803 undergraduate students in a questionnaire study of the effects of professor gender, student gender, and divisional affiliation on student ratings of professors and professor self-ratings. Students rated their professors on 26 questions tapping five teaching factors as well as overall teaching effectiveness. Professors rated themselves on the same questions as well as on nine exploratory ones. On student ratings, there were main effects for both professor gender (female professors were rated higher than male professors on the two interpersonal factors) and division (natural science courses were rated lowest on most factors). These patterns were qualified by significant interactions between professor gender and division. Although professor self-ratings varied by division, there were few significant correlations between professor self-ratings and students' ratings. Implications for future research are discussed.

Keywords Student ratings · College professors · Gender · Teaching · Divisional affiliation

Because of their importance for employment-related decisions, the validity of student ratings of college teaching has been of great concern. Despite research demonstrating validity through comparing academic performance of students in multiple-section courses on common examinations (higher-rated professors have students who perform better; d'Apollonia & Abrami, 1997), there have been troubling demonstrations of possible biasing factors, such as significant correlations of student ratings with expected course grade (but not actual grade) (Greenwald & Gillmore, 1997).

Appreciation goes to Laura Capotosto and Julie Phelan for their research assistance, and to all the faculty at Lafayette College that participated in this study.

Suzanne Montgomery is now at Widener Law School, Philadelphia, PA, USA.

S. A. Basow (✉) · S. Montgomery
Psychology Department, Lafayette College, Easton, PA 18042, USA
e-mail: basows@lafayette.edu

One potential source of bias is sexism: whether female faculty are evaluated more negatively than male faculty. Although research tends to find no gender difference on ratings of overall teaching effectiveness (Feldman, 1993), a number of studies suggest that professor gender may interact with other factors to produce lower ratings for women faculty. For example, studies by Basow (1995, 2000; Basow & Silberg, 1987) have found an interaction between professor gender and student gender such that male students often rate female faculty lower than female students do, and lower than male students rate male faculty. Other research has found similar interactions, often described as a same-sex preference, depending on the particular questions asked (Bachen et al., 1999; Centra & Gaubatz, 2000; Feldman, 1993; Hancock et al., 1993). For example, male and female students tend to rate same-sex teachers' vocal qualities higher than those of other-sex teachers.

In general, male faculty tend to be rated similarly by their male and female students while female faculty tend to be rated higher by female students and/or lower by male students. On questions relating to a faculty member's interactions with students, however, both sexes may give female faculty higher ratings than they give male faculty, although female students are likely to rate same-sex faculty the highest on such questions (Bachen et al., 1999; Basow, 1995; Centra & Gaubatz, 2000). For example, in Centra and Gaubatz's (2000) study of 741 classes at 21 institutions, male professors were evaluated similarly by their male and female students, but female professors were rated higher by their female students overall and on questions relating to communication and faculty–student interaction.

Interpretations of these findings focus on two possibilities: gender stereotypic expectations and gender-specific teaching styles. It is well documented that gender stereotypes lead to differential expectations of women and men (Biernat, 2003). For men, expectations regarding appropriate professional behavior overlap with expectations of masculinity: competence, dominance, high status, and authority. For women, however, there is only a narrow area of overlap between expectations of professional behavior and expectations of femininity. Professional women thus have to work harder to demonstrate feminine qualities of warmth, nurturance, and emotional sensitivity along with the aforementioned professional qualities. If they fail to meet this higher standard, they are judged more negatively. Research has found that it is more important for female professors to be friendly (to smile and be available) than male professors (Kierstead et al., 1988). It also is more important for female professors to be self-confident, stable, and steady (Burns-Glover & Veith, 1995). Women professors who give low grades or who do not have a warm expressive style are evaluated more negatively than their male counterparts (Sinclair & Kunda, 2000). Women faculty who lecture or who have a particularly informal style also may be evaluated more negatively than male colleagues demonstrating similar behaviors (Statham et al., 1991). Finally, women may have to be more available to students than their male colleagues in order to get similar ratings (Bennett, 1982).

Different ratings of male and female professors by male and female students may also be due to differential compatibility of teaching styles. Because of different sets of expectations as well as different socialization experiences, female faculty may indeed have different teaching styles than their male colleagues and students may prefer the style more consistent with their own gender. Several research studies report that female faculty tend to be more student-oriented and less authoritative than male faculty: they have more class discussion, less lecturing, and greater availability outside of class (Bennett, 1982; Centra & Gaubatz, 2000; Statham et al., 1991).

These teaching style differences may partially account for the higher ratings female faculty tend to receive on questions tapping interactions with students.

One potential confounding factor in studies of teaching style and student ratings is divisional affiliation. Although women currently are 38% of full-time faculty (Cataldi, Fihimia, & Bradburn, 2005), they are more likely to be found in humanities departments than in engineering and the physical sciences. Class discussions may be more likely in the humanities than in the sciences and engineering, so what looks like a gender difference may actually be a discipline difference. A common finding in the student evaluation literature is for male professors to be rated as more organized than female professors (Bachen et al., 1999; Basow, 1995, 2000; Basow & Silberg, 1987; Centra & Gaubatz, 2000). This could be a gendered behavior or a stereotypic perception, but it also could be due to the type of classes male professors teach. Natural science and engineering courses may lend themselves more to organized lectures than do humanities courses. Unfortunately, most studies do not control for divisional affiliation when examining gender dynamics.

Typically, courses in the natural sciences and engineering receive lower overall ratings than courses in the humanities (Basow, 1995; Marsh & Roche, 1997; Santhanam & Hicks, 2002), and particularly low ratings on faculty immediacy; that is, how close a professor is to a student (Moore et al., 1996). However, divisional affiliation may interact with professor gender and result in differential student ratings. Because women faculty are less likely to be in the physical sciences and engineering, their presence there may be viewed as particularly gender-inappropriate and they may be evaluated more negatively than are women in more gender-typical areas, such as English. Basow (1995) found an interaction between professor gender and divisional affiliation on several questions of an in-house student rating form, with the fewest difference in the humanities, supporting the idea that the humanities may be considered more gender-appropriate for females than other disciplines. In the same study, female professors were rated slightly lower than male professors on all questions in the natural sciences, perhaps because these courses are considered the most gender-inappropriate for females. In the social sciences, the pattern was mixed: female faculty received lower ratings on some questions but higher ratings on others, mainly those tapping faculty student interactions. Centra and Gaubatz (2000) also found professor gender and divisional affiliation to interact on certain questions in the social and natural sciences, but not in the humanities. In their study, female faculty in the natural sciences received higher ratings than their male counterparts on ratings of faculty-student interaction.

Thus any study of student ratings of professors must take into account professor gender, divisional affiliation, and student gender. We also need to examine a variety of teaching aspects, since different questions appear to show different patterns. Such an examination is the focus of the current study.

To examine the question of whether teaching style actually differs by professor gender or divisional affiliation, we asked professors to rate themselves on the same questions as the students in their classes did as well as on several other questions tapping aspects presumed to differ by gender and/or division. Statham et al. (1991) observed 167 professors in the classroom, examined their student evaluations, and interviewed 30 full-time professors matched for rank. They found that female professors were more student-oriented than male professors and were significantly more likely to encourage class participation and to actually have students participate in class and give presentations.

Based on previous research we expected to find gender and divisional patterns in both student ratings of professors and faculty self-ratings. With respect to students, Hypothesis 1 predicts a main effect of professor gender on interpersonal questions, with female faculty receiving higher ratings than male faculty (Bachen et al., 1999). Hypothesis 1a predicts significant interactions between professor gender and student gender (Basow & Silberg, 1987), especially on Scholarship, Organization/clarity, Dynamism/enthusiasm and overall, with each gender giving higher ratings to same-gender faculty. Hypothesis 2 predicts a main effect of course division, with humanities professors receiving the highest ratings and natural science professors the lowest. Hypothesis 2a predicts an interaction between professor gender and division, especially for overall teaching ability (Basow, 1995), with male professors receiving higher ratings than female professors in the natural sciences but not in the humanities.

With respect to professor self-ratings, Hypothesis 3 predicts no gender difference on the teaching factors but a difference on the exploratory questions, with female faculty reporting being more available to students and lecturing less compared to male colleagues. Hypothesis 4 predicts divisional differences on the exploratory questions as well, with faculty in the humanities reporting the most time spent in discussions and least time in lectures.

Finally, since previous research reports that student ratings and faculty self-ratings are significantly correlated (Feldman, 1989; Marsh, 1982), Hypothesis 5 predicts significant correlations on all five factors of a multidimensional rating form used by Basow and Silberg (1987).

Method

Participants

The study was conducted at a small liberal arts college in northeastern U.S. The sample was comprised of 43 professors, 23 males and 20 females (24% of the faculty overall, but 42% of the female faculty); and 803 students, 407 males, 365 females, and 31 students who did not report their gender (these students were excluded from analyses involving student gender). The students were 17 to 26 years old (M = 19.40, SD = 1.14) and were mostly in their first (33.8%) or second year (35.1%) of college.

Only professors teaching classes in the 100 or 200-level who had been teaching at the college at least 1 year and were at the rank of assistant professor or higher were requested to participate. Approximately 70% agreed. Only 100- and 200-level courses were used to increase the heterogeneity of the classes (with respect to gender and major) and to minimize the self-selection factor that occurs in upper level courses. The professors were generally matched by rank in three divisions: humanities, N = 17 (nine male, eight female), natural sciences, N = 18 (11 male, seven female), and social sciences, N = 8 (four male, four female). Overall, 27% were full professors, 22% associate professors, and 51% assistant professors. Because of the matching, the sample had a lower percentage of tenured professors (49%) than the population (67%). However, Chi square analyses indicated no significant differences between male and female professors by rank or division ($p > 0.9$). Class size ranged from 4 to 47, M = 21.8 (SD = 12.2).

Materials

The teacher rating form used by Basow and Silberg (1987) was used. The form (previously adapted by Leventhal et al., 1977 from Hildebrand & Wilson, 1970) had 26 questions concerning teaching and the students were asked to rate their professors on each one using a five-point Likert scale, where “1”=strongly disagree and “5” = strongly agree. The scale had five questions for each of the five factors: Scholarship (α = 0.61), Organization/clarity (α = 0.83), Instructor–Group Interaction (α = 0.82), Instructor–Individual Student Interaction (α = 0.86), and Dynamism/enthusiasm (α = 0.86). The 26th question concerned professors’ overall teaching effectiveness, a question that appears on virtually all student rating forms.

An adapted form of the Teacher Rating Form was used along with nine additional exploratory questions for professor self-ratings. Two of the additional questions are directly from the interview questions of Statham et al. (1991) . See Appendix. The questions were quantified using a five-point Likert scale to be consistent with the Teacher Rating Form.

Procedure

Over the course of two semesters, male and female professors who met the criteria were matched within their division by rank. The professors received an e-mail requesting their participation and those who did not respond were then contacted by a phone call, a visit, or another e-mail. Two-thirds of the 44 professors asked to participate in the first semester, and 74% of the 19 professors asked to participate in the second semester, agreed to do so.

A female student researcher distributed the questionnaires to each class and professor during the first 15 min of the class period, some time during the seventh through twelfth weeks of a 14-week semester. Student participation was voluntary and students who had already filled out the questionnaire in another class were asked not to fill it out again. After signing the informed consent forms, students were given their questionnaires to complete in the classroom and the professor was asked to leave the room to complete his or her questionnaire.

The questionnaires were collected and placed in an envelope that was coded for the division of the class, the gender of the professor and the rank of the professor. Therefore, the authors were unable to identify the name of the professor or the department from which the data came.

At the end of each semester, the professors were sent letters thanking them for their participation and were given debriefing statements to be distributed to their students.

Design

For student ratings of professors, a 2 (gender of student) × 2 (gender of professor) × 3 (division) between-factorial design was planned. For the professor self-ratings, a 2 (gender of professor) × 3 (division) between-factorial design was planned. We also examined the correlation between student evaluations of their professors on each of the five factors and the professor self-ratings on the same five factors plus the additional nine questions.

Results

Student Ratings

Male students, 52.7% of the sample, were under-represented among seniors (44.4%) and slightly over-represented among first year students (59.2%) (χ^2 (3) = 8.18, $p <$ 0.05). Since class year was significantly correlated with ratings of Organization (r (763) = −0.07, $p < 0.05$) (as class year increases, ratings of Organization/clarity decrease), it was used as a covariate in the planned MANOVAs. There was no gender difference in student age and age was not significantly correlated with any of the student ratings.

For student ratings of professors, a 2 (gender of professor) × 2 (gender of student) × 3 (division) multivariate analysis of covariance (with class year as a covariate) was performed on the average rating of each of the five factors on the teacher rating form and the overall measure of teaching effectiveness.

There was a multivariate main effect of professor gender, F (6,682) = 4.32, $p <$ 0.001. See Table 1 for Ms and SDs for each of the dependent variables. As Hypothesis 1 predicted, there were significant univariate main effects of professor gender on Instructor–Group Interaction F (1,687) = 7.44, $p < 0.01$ and Instructor–Individual Student Interaction F (1,687) = 4.25, $p < 0.05$. On both factors, female professors were rated higher than male professors.

Although Hypothesis 1 was supported, Hypothesis 1a, which predicted an interaction between professor gender and student gender, was not.

As Hypothesis 2 predicted, there was a multivariate main effect of course division, F (12,1366) = 8.94, $p < 0.001$. See Table 2 for Ms and SDs. The univariate effects of division were significant on Organization/clarity, F (2,687) = 4.21, $p < 0.05$, Instructor–Group Interaction, F (2,687) = 9.21, $p < 0.001$, Dynamism/enthusiasm, F (2, 687) = 6.28, $p < 0.01$, and overall teaching effectiveness, F (2, 687) = 5.10, $p < 0.01$. The main effect of course division on all four ratings is due to the significantly lower ratings of natural science professors. Tukey post hoc tests indicated that for Organization/clarity, professors teaching in the natural sciences were rated significantly lower than professors teaching in the social sciences. For Instructor–Group Interaction and overall teaching effectiveness, professors teaching in the natural sciences were rated significantly lower than professors teaching both in the humanities and the social sciences. For Dynamism/enthusiasm, professors teaching in the natural

Table 1 Mean student ratings and SDs of professors on five teaching factor and overall as a function of professor gender

	Male professors		Female professors	
	Mean	*SD*	Mean	*SD*
Scholarship*	3.46	0.64	3.58	0.98
Organization/clarity	3.64	0.72	3.67	0.78
Instructor–Group Interaction**	3.31	0.73	3.54	0.72
Instructor–Individual Interaction*	3.74	0.74	3.87	0.78
Dynamism/enthusiasm	3.94	0.79	4.00	0.82
Overall teaching effectiveness	3.71	0.84	3.81	0.92

*$p < 0.05$, **$p < 0.01$.

Table 2 Mean student ratings and SDs of professors on five teaching factors and overall as a function of division

Factor	Humanities		Natural sciences		Social Sciences	
	(*N* = 17)		(*N* = 18)		(*N* = 8)	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Scholarship	3.60	1.10	3.46	0.64	3.50	0.72
Organization/clarity*	3.66^{ab}	0.70	3.58^{b}	0.78	3.81^{a}	0.73
Instructor–Group Interaction***	3.53^{a}	0.73	3.29^{b}	0.72	3.54^{a}	0.73
Instructor–Individual Interaction	3.72	0.74	3.82	0.78	3.85	0.73
Dynamism/enthusiasm***	4.15^{a}	0.80	3.88^{b}	0.81	3.92^{b}	0.74
Overall teaching effectiveness**	3.88^{a}	0.81	3.64^{b}	0.90	3.85^{a}	0.88

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Means with different superscripts are significantly different.

sciences and social sciences both were rated significantly lower than professors teaching in the humanities.

Hypothesis 2a predicted that a professor gender x division interaction would qualify the main effects of both variables. This hypothesis was supported: F (12,1366) = 4.87, $p < 0.001$. Follow-up univariate tests were significant for Organization/clarity, F (2,687) = 5.12, $p < 0.01$, Instructor–Group Interaction, F (2,687) = 6.41, $p < 0.01$ and Dynamism/enthusiasm, F (2,687) = 3.26, $p < 0.05$. All three qualified significant main effects for division, and the Instructor–Group Interaction also qualified the main effect of professor gender.

Figure 1 depicts the Professor gender by Division interactions, with the results for Instructor–Group Interaction on top. Female professors were rated significantly higher than male professors in both humanities (M for female professors = 3.71, SD = 0.64; M for male professors = 3.37, SD = 0.78; t (230) = −3.62, $p < 0.001$) and natural science (M for female professors = 3.49, SD = 0.69; M for male professors = 3.16, SD = 0.71; t (377) = −4.59, $p < 0.001$) classes for Instructor–Group Interaction but significantly lower than male professors in social science classes (M for female professors = 3.44, SD = 0.77; M for male professors = 3.66, SD = 0.63; t (162) = 2.00, $p < 0.05$). The significantly lower ratings of natural science professors than professors in both other divisions, as revealed in the main effect of division on this variable, was due mainly to male professors. Female professors in the natural sciences were rated significantly lower than female humanities professors ($p < 0.05$), but not female social science professors.

The significant interaction for the Organization/clarity factor is shown in the middle of Fig. 1. Follow up t-tests show that there were no significant differences in how male and female professors were rated in the humanities and natural sciences on this factor, but in the social sciences, male professors (M = 4.00, SD = 0.56) were rated significantly higher than female professors (M = 3.64, SD = 0.79), t (164) = 3.46, p = 0.001, and significantly higher than male professors in the humanities (M = 3.57, SD = 0.69) and natural sciences (M = 3.54, SD = 0.74). Thus the main effect of division on this variable (highest ratings in the social sciences) was due entirely to ratings of male professors.

The follow up t-tests for the professor gender × division interaction on Dynamism/ enthusiasm (shown on bottom of Fig. 1) revealed no significant gender differences in the humanities or social sciences but in the natural sciences, female professors (M = 4.04, SD = 0.83) were rated significantly higher than were male professors (M = 3.79, SD = 0.77), contrary to predictions. Ratings for male professors were significantly higher in humanities courses (M = 4.10, SD=0.89) than in natural science courses (M = 3.79, SD = 0.77). For female professors, too, ratings were highest in humanities courses (M = 4.12, SD = 0.74), but this rating differed significantly only from ratings in social science courses (M = 3.83, SD = 0.81). Thus the

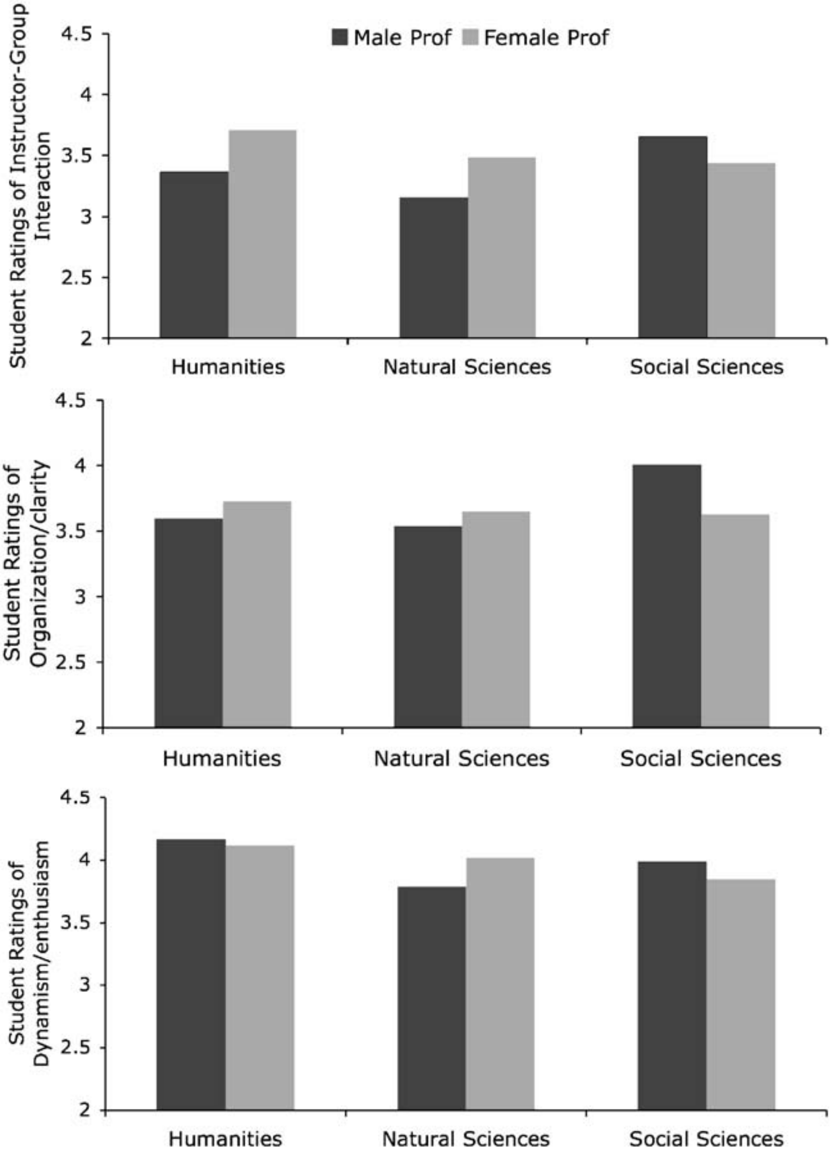

Fig. 1 Mean student ratings as a function of professor gender and division for Instructor-Group Interaction (*top*), Organization/clarity (*middle*), and Dynamism/enthusiasm (*bottom*)

significantly higher ratings of humanities courses than those in the other divisions (main effect of division) takes a different pattern for male and female professors: for male professors, Dynamism ratings were lowest in natural science courses but for female professors, Dynamism ratings were lowest in social science courses.

Student gender approached significance in the multivariate analysis, F (6,682) = 2.07, p = 0.055, but no univariate analyses were significant at p < 0.05. Female students gave slightly higher ratings than male students in general.

Professor Self-Ratings

Eleven professors skipped at least one question on the Self Rating Form. If a professor only answered three or four out of the five questions that make up each factor, the three or four responses were averaged so the professor would not have to be dropped from the sample. Data from two professors, one male and one female, both in the Humanities, had to be eliminated because these professors skipped so many questions on the rating form that it was impossible to compute factor scores. Therefore, the N for the Professor MANOVAs is 41. Seven professors neglected to answer the exploratory questions; the N for analyses of these questions is 34.

Cronbach's alpha was computed for each of the five factors on the professor-self rating form. All had good reliability, Scholarship, α = 0.76, Organization/clarity, α = 0.85, Instructor–Group Interaction, α = 0.81, Instructor–Individual Student Interaction, α = 0.86, Dynamism/enthusiasm, α = 0.90.

Because of small Ns, a two-way professor gender by division MANOVA was not possible; two one-way MANOVAs were run instead. As Hypothesis 3 predicted, there were no significant main effects of professor gender on the Teacher Rating Form but the effect of division approached significance, F (12, 68) = 1.77, p = 0.07, partially supporting Hypothesis 4. Because this hypothesis was exploratory, follow-up univariate analyses were conducted. Significant main effects of division were found for Instructor–Group Interaction, F (2, 38) = 8.51, p = 0.001, and overall teaching ability, F (2, 38) = 4.37, p < 0.05. Tukey post hoc tests indicated that humanities professors rated their Instructor–Group Interaction (M = 4.07, SD = 0.49) significantly higher than did social science (M = 3.36, SD = 0.74) and natural science professors (M = 3.21, SD = 0.66), the same pattern found in the student ratings. Humanities professors also rated their overall teaching ability (M = 4.0, SD = 0.38) significantly higher than did social science professors (M = 3.38, SD = 0.74), a pattern not found in the student ratings (where professors in the humanities were rated similarly to those in the social sciences but significantly higher than those in the natural sciences).

Responses to the nine exploratory questions were analyzed by professor gender using Chi square, collapsing some categories due to small cell sizes. For time spent per week, female professors were somewhat more likely than male professors to say they spend more than 3 h preparing (31.3% compared to 11.1%), and male professors were somewhat more likely than female professors to spend 1 h or less (33.3% compared to 6.3%) ($\chi^2(2)$ = 4.76, p < 0.10). There were no other significant gender differences in self ratings. Therefore, Hypothesis 3 was not supported. Most professors rated themselves as spending the same or more time preparing for class as their peers (82.3%), lecturing 60% of the time or more (67.6%), using discussion 20% of the time or more (52.9%), using group work less than 20% of the time (73.5%), having three or more student visits per week (70.6%), and being available to students outside of class 5 h or more per week (76.5%). Most also rated themselves as organized or very organized (82.4%), and considered student participation as important or very important (73.5%).

There were two differences by division, mainly on the percent of time spent lecturing, χ^2 (4) = 9.22, $p < 0.06$, and percent of time spent in discussion χ^2 (2) = 8.86, $p = 0.01$. As Hypothesis 4 predicted, most humanities professors (61.5%) spend less than 60% of time lecturing while 85.7% of natural science and social science professors spend 60% or more time lecturing. Similarly, most natural science (71.4%) and social science (57.1%) professors spend less than 20% of their time in discussion compared to only 15.4% of humanities professors. More than half of the latter (53.8%) spend 40% of the time or more in discussion.

Correlations

Classwise correlations (N = 41) were computed between the class average of students' ratings of their professor and their professor's self-ratings. See Table 3. In general, significant correlations between student ratings and professor self-ratings on the factor scores and overall were minimal, contrary to Hypothesis 5. Only student ratings of Dynamism/enthusiasm were significantly correlated with faculty self-ratings on that factor, r (41) = 0.384, p = 0.01, although student ratings of Scholarship were marginally correlated with faculty self-ratings on the same factor (p = 0.07). Student ratings of Organization were negatively correlated with faculty self-ratings of Instructor–Group Interaction, r (41) = −0.303, p = 0.05, and percent of discussion used, r (34) = −0.368, $p < 0.05$, and positively correlated with faculty self-ratings of organization, r (34) = 0.342, $p < 0.05$, and percent of lecturing used, r (34) = 0.39, $p < 0.05$, although not the Organization factor itself ($p > 0.10$). As

Table 3 Intercorrelations between student ratings of professors and professor self-ratings in same class on factor scores (N = 41) and selected exploratory questions (N = 34)

Professor Self-Ratings	Student ratings (factor scores and overall)					
	Scholarship	Organization	Instructor–Group Interaction	Instructor–Individual Student	Dynamism	Overall
Scholarship	0.285	−0.156	0.137	−0.019	0.157	0.033
Organization	−0.099	0.078	−0.170	−0.116	0.005	0.045
Instructor–Group Interaction	0.035	−0.303*	0.242	−0.223	0.060	−0.086
Instructor–Individual Student Interaction	−0.121	−0.134	−0.072	−0.156	−0.028	−.058
Dynamism	0.167	−0.001	0.122	0.089	0.384*	0.190
Overall	0.074	−0.006	0.045	−0.174	0.207	0.116
Organized Q	0.165	0.342*	−0.051	0.157	0.094	0.278
Lecture %	−0.086	0.390*	−0.254	0.132	−0.129	0.042
Discussion %	−0.030	−0.368*	0.162	−0.293	−0.025	−0.127

*$p < 0.05$.

professor self-ratings of organization and percent of time spent lecturing increased, and self-ratings of Instructor–Group Interaction and percent of time spent in discussion decreased, student ratings of their professor's Organization increased. Student ratings of Instructor–Group Interaction, Instructor–Individual Student Interaction, and overall ratings were not correlated significantly with any faculty self-ratings.

Discussion

Consistent with the major hypothesis, professor gender operated in complex ways in student ratings of professors, varying by division and the particular teaching aspect being rated. Thus it is critical in student evaluation research to examine or control for these variables. The five specific hypotheses, however, received only mixed support. The most surprising finding was the lack of much correspondence between student ratings of faculty and faculty self-ratings.

As Hypothesis 1 predicted, female professors received higher ratings than male professors on the interpersonal questions, Instructor–Group Interaction and Instructor–Individual Student Interaction (Bachen et al., 1999; Bennett, 1982; Centra & Gaubatz, 2000). However, the higher ratings of female professors on Instructor–Group Interaction only occurred in humanities and natural science courses; the pattern was reversed in social science courses (significant professor gender by division interaction). Since there were only eight social science professors in the sample (half female), the social science results may be idiosyncratic and should be interpreted cautiously. As previous research suggests, female faculty may be more student-oriented than male faculty and convey that orientation both on a one-to-one level and in classroom dynamics (e.g., Statham et al., 1991). Centra and Gaubatz's (2000) multi-college study attributed the higher ratings of female professors on questions tapping faculty–student interactions to the tendency of female faculty to use more discussion and less lecturing. In the present study, however, professor self-ratings indicated no significant gender differences on the two interpersonal factors in question, nor in ratings of the importance of student participation, time spent lecturing, discussion or group work, or time available outside of class or number of student visits. Thus student-perceived gender differences in this study may reflect subtle qualitative differences in male and female professors' interactions with their students, or student ratings may reflect students' gender stereotypes more than professors' actual behaviors (Biernat, 2003).

The lack of gender differences in professors' self-ratings of these interpersonal behaviors was contrary to Hypothesis 3 but may be due to the fact that the present study was conducted at a small private liberal arts campus where faculty–student interaction is highly valued. Perhaps at larger institutions where class size is larger and/or teaching is less emphasized (such as where Statham et al., 1991, conducted their study and where half of Centra & Gaubatz's sample was obtained), more gender differences might be found. The lack of professor gender differences in student-related behaviors in the present study may also be due to the fact that professors were matched for division. Since female faculty tend to be over-represented in humanities divisions relative to the physical sciences and engineering, and humanities professors are generally rated highest in Instructor–Group Interaction by both students and themselves, it may be that divisional differences appear as gender differences when division is not controlled for. The present study, in which 53% of the humanities professors were male, controlled for that potential problem.

Indeed, divisional differences were significant in both faculty and student ratings, supporting Hypotheses 2 and 4. Humanities professors indicated that they spent significantly more class time in discussion than did natural and social science professors, who spent more class time lecturing. Students too noted divisional differences, rating humanities professors higher than natural science professors in Instructor–Group Interaction as well as in Dynamism/enthusiasm and overall teaching ability. Social science professors were rated significantly higher than natural science professors in Organization/clarity, Instructor–Group Interaction, and overall. These results support previous findings that student ratings vary by division, with humanities courses frequently receiving the highest ratings and natural science courses frequently receiving the lowest (Basow, 1995; Marsh & Roche, 1997; Moore et al., 1996; Santhanam & Hicks, 2002). It might be that the lower ratings of natural science professors by students on Instructor–Group Interaction may be directly related to the greater time faculty in that division spend lecturing. However, there is no significant correlation between student ratings on this variable and faculty self-ratings of time spent lecturing.

Although natural science professors typically received the lowest ratings, three out of four divisional differences were qualified by interactions with professor gender. The lower ratings of natural science professors appear more in ratings of male than female faculty, contrary to predictions (Hypothesis 2a). Male faculty also were rated significantly lower than female faculty in the natural sciences on ratings of Instructor–Group Interaction as well as Dynamism/enthusiasm. It may be that female professors in the natural sciences, at least in this sample, take extra steps to engage students and appear enthusiastic in order to counter negative gender stereotypes of women in nontraditional fields. Similar results were found for ratings of faculty–student interactions in Centra and Gaubatz's (2000) multi-institutional study. The small numbers in the faculty self-ratings in the present study did not allow for analysis by division and gender, but other research suggests that women professors, regardless of field, are more student-oriented and use discussion more than male professors (Centra & Gaubatz, 2000; Statham et al., 1991). Because of the smaller percentage of women in most science fields, what typically appears to be a divisional main effect in other studies may actually mask this gender pattern.

The higher ratings of female compared to male natural science professors (on Instructor–Group Interaction and Dynamism/enthusiasm) contradict the findings of Basow and Silberg (1987) at the same school using the same student rating form. In that study, both male and female students rated male faculty higher than female faculty in the natural sciences. It is possible that the difference in results between the studies is due to the increase in the number of women in the natural sciences in the 16 years that separate the two studies, thus making the sciences less a gender-incongruent field for women. In addition, it also is possible that the current women faculty in the natural sciences are more proficient in their teaching than those in the past. Due to the small number of female professors sampled in the natural sciences ($N = 7$), it may also be that one or two of the professors examined were exceptional teachers. Another possibility for the lack of negative ratings for female natural science professors may be due to changes in student attitudes over the 16 years that separate this study from that of Basow and Silberg. The student body may have become less sexist and more accepting of women in nontraditional fields.

Support for the latter hypothesis is found in the fact that student gender had little effect on the results in the current study, contrary to predictions (Hypothesis 1a). Overall, students appear to rate faculty similarly regardless of their own gender. These results have occasionally been found in other research. Indeed, many researchers have remarked on the conflicting results found when examining gender variables as well as the small effect size of such interactions when they are found (Centra & Gaubatz, 2000; Feldman, 1993). Perhaps the matching of professors on rank and division in the present study minimized the effects of gender. Rank, for example, is related negatively to student ratings of Dynamism/enthusiasm (r (787) = -0.179, $p < 0.05$) and overall teaching ability (r (803) = -0.118, $p < 0.05$) in the present study. Since female faculty generally are in lower ranks in academia than male faculty, some of the inconsistencies in the literature may be due to not controlling for this variable.

Perhaps the most surprising finding in the study was the minimal correlation between student ratings and faculty self-ratings, contrary to Hypothesis 5 and previous research (Feldman, 1989; Marsh, 1982). Even though the rating instrument had not previously been used for faculty self ratings, the reliability of each of the five factors was strong (alphas = 0.76–0.90). Furthermore, the pattern of faculty results somewhat paralleled those of students with respect to divisional differences on ratings of Instructor–Group Interaction and overall teaching ability (humanities professors were high, natural science professors low, supporting Hypothesis 4). However, there were no significant correlations between student and faculty ratings of Instructor–Group Interaction, Instructor–Individual Student Interaction, or overall ratings. Since all three ratings were qualified in the student sample by two-way interactions, it may be that similar patterns might emerge in the faculty data if we had enough faculty participating to examine interactions. More faculty also would have given the analyses more power, especially given Feldman's (1989) finding based on 19 studies that the average correlation between faculty self-ratings and current student ratings is only 0.29.

Correlations between student ratings and faculty self-ratings on the same questions were significant only for Dynamism/enthusiasm and marginally for Scholarship. Although student ratings of Organization/clarity were not significantly correlated with faculty self-ratings on this factor, student ratings were significantly correlated with the single faculty question on the same topic, indicating some construct validity. Given the other significant correlates of the student ratings of Organization (high faculty self-ratings of Instructor–Group Interaction and percent of discussion used, and low faculty self-ratings of percent of lecturing used), the factor may measure perceptions related to group discussions more for students than it does for faculty. This should be kept in mind when trying to interpret student ratings on this factor. Feldman (1989) also found lower faculty–student correlations on dimensions relating to teacher preparation and organization of the course than on dimensions relating to teacher stimulation of interest (similar to Dynamism/enthusiasm).

The lack of more congruence between student and faculty ratings is particularly surprising because at the college in question, student evaluations are done on every course in every semester. Since all the faculty in the sample were at least in their second year of teaching at the college, professors should have a sense of what students think of their teaching and could use that sense to evaluate themselves. This does not seem to be the case. This suggests that, with respect to some questions at least, either faculty disagree with their students' perceptions or that they interpret the questions differently than their students. It may also be that students misperceive faculty intentions or misinterpret faculty behavior. Whatever the case, the lack of congruence means there is a perceptual gap that needs to be overcome in order for faculty to understand the meaning of student ratings. If students and faculty are referring to different behaviors when evaluating teaching effectiveness, then student feedback will not help faculty members improve their teaching. Indeed, faculty members may not understand why students rate them the way they do. This misunderstanding may contribute to a real or perceived disconnection between students and faculty members. It does not seem the case that faculty only are rating themselves positively since the average faculty rating for both teachers and students are between a “3” and “4” on a five-point scale.

A major limitation of the study is that there were not enough professors to examine professor gender and divisional effects on faculty self-ratings, nor enough power to find differences even when examining main effects. In addition, the size of the classes studied varied from 4 to 47 students. This difference in class size may have had an impact on the results, especially those relating to ratings of interpersonal factors, although class size did not differ significantly as a function of professor gender or divisional affiliation. A related problem is that the professors who agreed to participate may be the strongest or most confident professors, and the results therefore might not generalize to the population of professors. However, the response rate was fairly high (66–74%), so it is likely that the results represent at least this population, although perhaps not professors at dissimilar institutions, such as major research universities. In the same vein, the majority of both students and professors were White, so results may not generalize to other race and ethnicities.

Another limitation of this study was that verbal rather than behavioral measures were used, thus measuring what professors and students say professors do in class. It would be interesting to see what professors actually do in these classes and how students perceive individual teaching behaviors, and whether either of these vary with professor and/or student gender. Certainly, the lack of agreement between student ratings and faculty self-ratings of the same questions needs further exploration, especially since student ratings often are the main (or sole) criterion of teaching effectiveness when personnel decisions are made.

Overall, student ratings of faculty are affected by many factors: professor gender, divisional affiliation, and the specific questions asked. We need more studies of interactional effects using multidimensional student ratings forms in order to properly interpret and utilize student ratings. Given their importance in the career trajectory of most faculty members, such research is imperative.

Appendix

Additional Questions:

Please answer the following questions about this particular class by circling the appropriate answer.

1. How much time do you spend preparing for a class period?
   Less than 1 h 1h 2–3 h 4–6 8 h or more

2. Do you think you spend as much time/less time or more than your colleagues at your rank?
   Much less   less   the same   more   much more
3. How organized would you say your classes are?
   Not organized   somewhat organized   neutral organized   very organized
4. What percentage of a typical class do you spend on lecture?
   Less than 20%   20–39%   40–59%   60–79%   80% or more
5. What percentage of a typical class do you spend on discussion?
   Less than 20%   20–39%   40–59%   60–79%   80% or more
6. What percentage of a typical class do you spend on student presentations and small group work?
   Less than 20%   20–39%   40–59%   60–79%   >80% or more
7. Part of the teaching role involves seeing students outside of class in your office. About how many students come to your office each week?
   0–1   1–2   3–4   5–6   7 or more
8. How many hours per week on average are you available to students outside of class?
   Less than 1 h   1–2 h   3–4   5–6 h   more than 6 h
9. How important is class participation?
   Very unimportant   Somewhat unimportant   Neither important nor unimportant   Important   Very Important

Note: adapted from *Gender and University Teaching* (pp. 172–176) by Statham et al. 1991, Albany: State University of New York Press. Copyright 1991 by State University of New York.

References

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. (2000). Best and worst professors: gender patterns in student choices. *Sex Roles 43*, 407–417.

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.

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

Burns-Glover, A. L., & Veith, D. J. (1995). Revisiting gender and teaching evaluations: sex still males a difference. *Gender in the Workplace 10*, 69–80.

Cataldi, E. F., Fahimi, M., & Bradburn, E. M. (2005). *2004 national study of postsecondary faculty (NSOPF: 04) report on faculty and instructional staff in Fall 2003.* Retrieved May 25, 2006, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005172.

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

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist 52*, 1198–1208.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (Neutral) observers. *Research in Higher Education 30*, 137–194.

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.

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

Hancock, G. R., Shannon, D. M., & Trentham, L. L. (1993). Student and teacher gender in ratings of university faculty: results from five colleges of study. *Journal of Personnel Evaluation in Education 6*, 235–248.

Hildebrand, M., & Wilson, R. (1970). *Effective University Teaching and its Evaluation*. Washington, District of Columbia: U.S. Department of Health, Education, and Welfare, Office of Education.

Kierstead, D., D'Agostino, P. D., & Dill, H. (1988). Sex role stereotyping of college professors: bias in students' ratings of instructors. *Journal of Educational Psychology 80*, 342–344.

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

Marsh, H. W. (1982). Validity of students' evaluations of college teaching: a multitrait-multimethod analysis. *Journal of Educational Psychology 74*, 264–279.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist 52*, 1187–1197.

Moore, A., Masterson, J. T., Christophel, D. M., & Shea, K. A. (1996). Making students' evaluations of teaching effectiveness effective. *Communication Education 45*, 29–39.

Santhanam, E. & Hicks, O. (2002). Disciplinary, gender and course year influences on student perceptions of teaching: explorations and implications. *Teaching in Higher Education 7*, 17–31.

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

Statham, A., Richardson, L., & Cook, J. A. (1991). *Gender and University Teaching*. Albany, New York: State University of New York.