Scientific
Research
Publishing

A Gendered Study of Student Ratings of Instruction

Lucas Huebner, Rhonda C. Magel

Department of Statistics, North Dakota State University, Fargo, ND, USA
Email: Rhonda.magel@ndsu.edu

Abstract

This research tests for differences in mean class averages between male and female faculty for questions on a student rating of instruction form at one university in the Midwest are considered to be in the category of "very high research activity" by the Carnegie Commission on Higher Education. Differences in variances of class averages are also examined for male and female faculty. Tests are conducted by first considering all classes across the entire university and then classes just within the College of Science and Mathematics. The proportion of classes taught by female instructors in which the average male student rating was higher than the average female student rating was compared to the proportion of classes taught by male instructors in which the average male student rating was higher than the average female student rating. Results are discussed.

Keywords

Gendered Expectations, Means, Variances, Proportions

1. Introduction

Student ratings of instruction are often used by universities as the main way of evaluating the teaching effectiveness of a faculty member. This is particularly true at research universities [1]. Some studies have shown that there is gender bias in student ratings of instruction [2] [3]. Marcotte [4] discusses a small study conducted based on student ratings in online courses. When students thought the instructor was female, the instructor was rated lower in all 12 categories considered. This included the category of "caring and respectful". Everything about the online course was the same in all cases except for students were told different genders for the instructor. One category in the study discussed by Marcotte [4] is "promptness" of the instructor in returning graded items. Students who thought the instructor was male gave the instructor an average rating of 4.35. Students who thought the instructor was female, gave the instructor an average rating of 3.55.

In other research, Feldman [5] [6] and Lueck, Endres, & Caplan [7], found that female students rated female instructors higher and male students rated male instructors higher. Feldman [5] [6] found that this was evidenced further based one's perception of gender "roles", which is more pronounced among business and engineering students. Worthington [8], in particular found this gender bias among finance majors. Basow [9] and Centra & Gaubatz [10] found that male instructors were rated similarly by both their male and female students. These studies found however, that female students tended to rate female instructors higher overall. Male instructors tend to be rated higher in knowledge/scholarship questions as well as enthusiasm. Female instructors tend to be rated higher in "comfortable environment" [11]-[13].

The subject area that an instructor teaches in also has an effect on student ratings of instruction. Teachers in science and engineering get lower ratings than those teachers in the humanities and social sciences. Female faculty fair even worse in the sciences, finding both male and female students rated female instructors lower than male instructors [9] [11] [12].

The age of an instructor was found to also affect ratings. Older instructors do not receive as high of student ratings of instruction. Increasing age has a more negative effect on female instructors [14].

Sprague and Massoni [15] and Laube, Massoni, Sprague, and Ferber [16] found that students have gendered expectations of instructors. Students expect female instructors to be caring and nurturing. They also expect female instructors to be available more often than male instructors in addition to not being as demanding or grading as harshly [12] [13] [15] [17] [18]. Students expect male instructors to be entertaining and energetic. Schmidt created a database on the words students used on the website "Rate My Professor" [19]. He found that patterns in the words used were associated with the gender of the instructor. The words "intelligent", "genius", and "smart" came up more often when evaluating male instructors across all disciplines. The words "bossy", "nurturing", and "strict" came up more when evaluating female instructors across all disciplines [19]. Researchers found that female instructors who did not meet these "gendered" expectations were rated more harshly than the male instructors who did not meet their gendered expectations [12] [13] [15] [17].

Studies have also shown that there is a significant positive correlation between the grade a student expects in a class and the same student's teaching evaluation of the professor [20]-[23]. In classes that have a common final exam, it has been shown that there is a modest significant positive correlation between student performance on the exam and the corresponding evaluation by the student of the teaching effectiveness of the instructor. However, based on the results of another study considering classes with no common final, there is some evidence that a student may give a higher rating to an instructor because they are "grateful" for the grade they are expecting to receive, or may give a lower rating to an instructor because they are upset about the grade they are receiving [24]. This could result in grade inflation [24]. Recall that researchers have found that female faculty are expected to grade less harshly than male faculty and are rated lower by students if they do not meet this gendered expectation [12] [13] [15] [17].

Basow and Martin [25] summarize much of the research that has been done concerning gendered expectations of students and completing student ratings of instruction for male and female faculty. They say that female faculty must be more caring and nurturing, be available more often, and not be as demanding on tests and assignments as male faculty, in order to get comparable evaluations. Female faculty not having these "gendered" characteristics get lower ratings. Findings from a North Carolina University study with lead researcher Lillian MacNell, suggest that female faculty still have to work harder to get similar ratings to male faculty members [4].

2. Study Design

We would like to explore some of these gendered findings at an intensive research university in the Midwest to see if we find similar results. In particular, we would like to examine student ratings of instruction for the 2013-14 academic year at North Dakota State University. This university is located in Fargo, North Dakota. It is ranked by the Carnegie Commission on Higher Education as one of the top public and private universities in the country and by NSF in the category of "Research University/Very High Research Activity". As of 2014, the university had a total enrollment of 14,747 students broken down into 12,124 undergraduate students, 340 professional students, and 2283 graduate students. Approximately 91% of the students were US citizens, 7% international, and 2% permanent residents with 54% of the students being male and 46% being female. The University consists of 8 colleges (www.ndsu.edu). Institutional Review Board (IRB) approval was obtained.

At North Dakota State University (NDSU), all instructors are required to give students in their classes an op-

portunity to evaluate the instruction in the class. This evaluation must take place during the last three weeks of the semester. The student rating of instruction form used at NDSU currently consists of 16 questions. These questions are given in **Figure 1**. The first six questions on the form have been used during the past 10 years. The

Figure 1. Student rating of instruction form.

last ten questions on the form were added to the form in 2013. The form also asks students to respond to the following demographic questions: 1) Their gender; 2) Their level (freshman, sophomore, junior, senior, graduate student); 3) Whether the course is elective or not; and 4) Their expected grade (A, B, C, D, F).

Student response data for the academic year 2013-14 was collected from all classes taught in an in-class environment, but not including workshops, seminars, or independent study classes. Data was collected from a total of 2092 classes.

Three areas of research will be examined in regard to this student rating of instruction questionnaire. The first area that will be examined is how the average student response to each of the questions is related to the class demographics. In particular, the demographics that will be considered include the following: 1) The percentage of students in the class required to take the class for their major; 2) The percentage of males in the class; 3) The percentage of students expecting to receive either an A or B in the class; 4) The percentage of freshman and sophomores in the class; and then 5) The gender of the instructor. In this phase of the research, a total of 16 least squares regressions will be conducted with the dependent variables for each of the regressions being the average class responses to each of the questions. The independent variables for each of the regression models will be the five demographic variables that have been mentioned. We would like to determine how much of the variation in class average responses to each question is explained by the class demographics. If this percentage is high, this indicates that the question is not measuring effective instruction, but rather something else. In this phase of the research, we would also like to investigate whether or not the gender variable is significant.

The second area of research that we will examine is to compare the average class responses of each of the questions for female instructors with the average class responses of each of the questions for male instructors across all 8 colleges of the university. We will test to see if there is a significant difference in the average class responses between male and female instructors for each question using two versions of a t-test: one assuming equal variances (pooled); and one not assuming equal variances (Satterthwaite). We are particularly interested in how the means compare between male and female instructors for the following questions: Question 5, "The fairness of procedures for grading this course", and Question 10, "I understood how my grades were assigned in this course", since research indicates that students expect female instructors to grade them less harshly [12] [13] [15] [17]; Question 12, "The instructor was available to assist students outside of class", because research has shown that students expect female instructors to be available more often [12] [13] [15] [17]; and Question 13, "The instructor provided feedback in a timely manner", since research has shown students rate male instructors higher in this category even if response time is the same [4]. The average class responses of each questions between male and female instructors within the College of Science and Mathematics will also be compared. Research has shown that female instructors in the sciences tend to get lower ratings than male instructors from the students [9] [11] [12]. In addition to comparing the mean class average responses between males and females for both the entire University and then for the College of Science and Mathematics, the variances in the average class responses between male and female instructors will also be compared for each question considering the entire University and then considering only the College of Science and Mathematics.

In the last phase of the research, we will use only classes in which at least five male students and five female students responded. In this phase we will compare the proportion of classes taught by female instructors in which the average male student response is higher than the average female student response to the proportion of classes taught by male instructors in which the average male student response is higher. These proportions will be compared for each question. Our hypothesis is that the proportions will be higher for male instructors. Our sample consisted of 112 classes taught by female instructors and 162 classes taught by male instructors. The lower sample size was due to the lack of students not responding to the demographic question about their gender and to the fact that we deleted any class from this portion of the study that did not have at least 5 male and 5 female students responding.

3. Results

Recall that in the first phase of the research, 16 ordinary least square regressions were to be conducted with the 16 dependent variables being the average class responses to each of the 16 questions. The independent variables in the model were the five class demographic variables collected: 1) The percentage of students in the class required to take the class for their major; 2) The percentage of males in the class; 3) The percentage of students expecting to receive either an A or B in the class; 4) The percentage of freshman and sophomores in the class;

and then 5. The gender of the instructor. For the majority of the questions, the demographics explained between 15% and 28% of the variation in responses. There were four questions that were the exception to this. Fifty-nine percent of the variation of in class average responses to question 11, "I met or exceeded the course objectives given for this course", were explained by the demographics. Forty-nine percent of the variation in class average response to question 6, "Your understanding of the course content"; 39% of the variation in class average responses to question 5, "The fairness of procedures for grading this course"; and 32% of the variation in class average responses to question 4, "The quality of this course", was explained by the demographics. For all questions, the percentage of students expecting to receive an A or a B was highly significant in explaining the variation in class average responses. For questions 7, 12, 15, and 16, the percentage of males in the class was significant at alpha equal to 0.05, and for question 8, the percentage of males in the class was significant at 0.10. In all of these cases, an increase of 10% males in the class corresponded with an average increase of approximately 0.04 in the instructor rating for that question. The only other demographic variable found to be significant for any question was the percentage of students taking the class because it was required for their major. This demographic variable was only significant (alpha = 0.05) for question 6. In this case, the class average response decreased with an increase in the proportion of students for which this course was required for their major. The first variable to enter any of the models, and the most significant variable, was the variable for the proportion of students in the class expected to receive either an A or a B grade. The indicator variable for gender of the instructor was not significant for any of these models with the proportion of students expecting an A or a B already in the model.

If this set of questions were used to evaluate effective teaching, one may consider dropping questions 6 and 11 since about 50% or more of the variation in class average responses is explained by the class demographics. These questions are not really evaluating effective teaching.

We next tested whether there was a difference in the mean class average responses for each of the 16 questions between classes taught by female instructors and classes taught by male instructors. The sample means for male and female instructors for each of the questions is given in **Table 1**.

When the classes across all 8 colleges of the university was considered, a significant difference was found in the mean class average responses between male and female instructors for questions 6 and 11 with alpha equal to 0.05 and for question 15 with an alpha value of 0.10 (p-value = 0.0548). In all these cases, the mean response for female instructors was higher. Female instructors were rated higher on the student's understanding of the course content, the student's meeting or exceeding course objectives, and the instructor setting and maintaining higher standards. It is interesting to note that when a regression analysis was conducted with class average response to question 11 being the dependent variable, gender of the instructor was not significant when the percentage of students who expected to get an A or a B was in the model. This was also true for questions 6 and 15. Students in female instructor classes were expecting more A's and B's than student's in male instructor classes.

Female instructors were not rated significantly lower on average to Question 5, "The fairness of grading procedures", but the sample mean of the class averages for female instructors was lower (p-value = 0.1242). Female instructors were not rated significantly lower on average to Question 10, "I understood how my grades were assigned", but the sample mean of the class averages for female instructors was slightly lower for this question (p-value = 0.6359). There was also no significant difference in means for Questions 12 and 13, about the availability of the instructor and providing feedback in a timely manner. The sample means for male and female instructors were very close to each other, with the sample means for females only being very slightly higher (p-values = 0.6848 and 0.8016, respectively).

Since research has shown that female faculty are rated lower in the science field [9] [11] [12], mean class average responses between male and female instructors were compared in the College of Science and Mathematics. These are given in **Table 2**. A significant difference at alpha equal to 0.05 was found between the mean responses for questions 3, 5, 7. A marginally significant difference was found between the mean responses of question 1 (p-value = 0.1004), and question 2 (p-value = 0.0824). In the three significant cases, except for question 5, the mean response associated with female faculty was higher. Female instructors were rated higher on communication ability, and creating an atmosphere conducive to learning. Research has shown that female faculty tend to be rated higher in "comfortable environment", but it is interesting to note we did not find a significant difference for the overall University, but in the College of Science and Mathematics for this question [11]-[13]. Females were rated marginally higher on the student's satisfaction with the instruction in the course, and on the instructor as a teacher. Question 5 had students rating the fairness of procedures for grading this course. Research

Table 1. Mean gender results for Questions 1 - 16 all colleges. (a) Mean gender results Question 1; (b) Mean gender results Question 2; (c) Mean gender results Question 3; (d) Mean gender results Question 4; (e) Mean gender results Question 5; (f) Mean gender results Question 6; (g) Mean gender results Question 7; (h) Mean gender results Question 8; (i) Mean gender results Question 9; (j) Mean gender results Question 10; (k) Mean gender results Question 11; (l) Mean gender results Question 12; (m) Mean gender results Question 13; (n) Mean gender results Question 14; (o) Mean gender results Question 15; (p) Mean gender results Question 16.

(a)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.2703	0.6180	0.0208	1.0000	5.0000
M	1199	4.2758	0.6191	0.0179	1.0000	5.0000
Diff (1-2)		−0.00552	0.6186	0.0274		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	−0.20	0.8405
Satterthwaite	Unequal	1910.9	−0.20	0.8404

P-value = 0.9570 when testing equality of variances.

(b)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.3301	0.6187	0.0208	1.0000	5.0000
M	1199	4.3420	0.6188	0.0179	1.0000	5.0000
Diff (1-2)		−0.0119	0.6188	0.0274		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	−0.44	0.6628
Satterthwaite	Unequal	1909.3	−0.44	0.6628

P-value = 0.9982 when testing equality of variances.

(c)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.2936	0.6469	0.0217	1.2500	5.0000
M	1199	4.2472	0.6710	0.0194	1.0000	5.0000
Diff (1-2)		0.0464	0.6609	0.0293		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	1.59	0.1127
Satterthwaite	Unequal	1946	1.60	0.1108

P-value = 0.2405 when testing equality of variances.

(d)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.1927	0.5889	0.0198	1.6000	5.0000
M	1199	4.2207	0.5878	0.0170	1.0000	5.0000
Diff (1-2)		−0.0280	0.5882	0.0261		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	−1.07	0.2831
Satterthwaite	Unequal	1907	−1.07	0.2832

P-value = 0.9505 when testing equality of variances.

(e)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.3222	0.5981	0.0201	1.0000	5.0000
M	1199	4.3617	0.5553	0.0160	1.0000	5.0000
Diff (1-2)		−0.0395	0.5739	0.0254		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	−1.56	0.1200
Satterthwaite	Unequal	1826.9	−1.54	0.1242

P-value = 0.9505 when testing equality of variances (females higher).

(f)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.2371	0.5477	0.0184	1.0000	5.0000
M	1199	4.1667	0.5614	0.0162	1.0000	5.0000
Diff (1-2)		0.0704	0.5556	0.0246		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	2.86	0.0043
Satterthwaite	Unequal	1934.4	2.87	0.0041

P-value = 0.4327 when testing equality of variances.

(g)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.3359	0.5629	0.0189	2.0000	5.0000
M	1199	4.3250	0.5547	0.0160	1.0000	5.0000
Diff (1-2)		0.0109	0.5582	0.0247		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	0.44	0.6586
Satterthwaite	Unequal	1893.5	0.44	0.6593

P-value = 0.6370 when testing equality of variances.

(h)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.2859	0.5621	0.0189	1.0000	5.0000
M	1199	4.2671	0.5698	0.0165	1.0000	5.0000
Diff (1-2)		0.0188	0.5665	0.0251		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	0.75	0.4529
Satterthwaite	Unequal	1923.1	0.75	0.4520

P-value = 0.6665 when testing equality of variances.

(i)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.2319	0.6026	0.0202	1.2500	5.0000
M	1199	4.2166	0.6196	0.0179	1.0000	5.0000
Diff (1-2)		0.0153	0.6124	0.0271		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	0.56	0.5739
Satterthwaite	Unequal	1937.3	0.56	0.5723

P-value = 0.3768 when testing equality of variances.

(j)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.3324	0.5664	0.0190	1.0000	5.0000
M	1199	4.3441	0.5415	0.0156	1.0000	5.0000
Diff (1-2)		−0.0117	0.5522	0.0245		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	−0.48	0.6336
Satterthwaite	Unequal	1860.1	−0.47	0.6359

P-value = 0.1489 when testing equality of variances.

(k)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.2201	0.5119	0.0172	1.0000	5.0000
M	1199	4.1705	0.5399	0.0156	1.0000	5.0000
Diff (1-2)		0.0496	0.5281	0.0234		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	2.12	0.0340
Satterthwaite	Unequal	1961.7	2.14	0.0327

P-value = 0.0906 when testing equality of variances.

(l)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.3247	0.5343	0.0179	1.7500	5.0000
M	1199	4.3149	0.5582	0.0161	1.0000	5.0000
Diff (1-2)		0.00979	0.5482	0.0243		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	0.40	0.6867
Satterthwaite	Unequal	1952.8	0.41	0.6848

P-value = 0.1643 when testing equality of variances.

(m)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.2960	0.5839	0.0196	1.2500	5.0000
M	1199	4.2896	0.5715	0.0165	1.0000	5.0000
Diff (1-2)		0.00644	0.5768	0.0255		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	0.25	0.8009
Satterthwaite	Unequal	1886.2	0.25	0.8016

P-value = 0.4901 when testing equality of variances.

(n)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.2817	0.5706	0.0192	1.6000	5.0000
M	1199	4.2610	0.5718	0.0165	1.0000	5.0000
Diff (1-2)		0.0207	0.5713	0.0253		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	0.82	0.4132
Satterthwaite	Unequal	1911.4	0.82	0.4131

P-value = 0.9489 when testing equality of variances.

(o)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.3334	0.4834	0.0162	2.0000	5.0000
M	1199	4.2909	0.5212	0.0151	1.0000	5.0000
Diff (1-2)		0.0425	0.5054	0.0224		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	1.90	0.0576
Satterthwaite	Unequal	1981.4	1.92	0.0548

P-value = 0.0168 when testing equality of variances (males higher).

(p)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	887	4.2866	0.5496	0.0185	1.0000	5.0000
M	1199	4.2998	0.5350	0.0155	1.0000	5.0000
Diff (1-2)		−0.0133	0.5413	0.0240		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	2084	−0.55	0.5804
Satterthwaite	Unequal	1880.2	−0.55	0.5819

P-value = 0.3904 when testing equality of variances.

Table 2. Mean gender results for questions 1 - 16 college of science and math. (a) Mean gender results Question 1; (b) Mean gender results Question 2; (c) Mean gender results Question 3; (d) Mean gender results Question 4; (e) Mean gender results Question 5; (f) Mean gender results Question 6; (g) Mean gender results Question 7; (h) Mean gender results Question 8; (i) Mean gender results Question 9; (j) Mean gender results Question 10; (k) Mean gender results Question 11; (l) Mean gender results Question 12; (m) Mean gender results Question 13; (n) Mean gender results Question 14; (o) Mean gender results Question 15; (p) Mean gender results Question 16.

(a)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.2393	0.5589	0.0420	2.3717	5.0000
M	359	4.1506	0.6387	0.0337	1.0000	5.0000
Diff (1-2)		0.0887	0.6135	0.0563		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	1.57	0.1161
Satterthwaite	Unequal	395.07	1.65	0.1004

P-value = 0.0450 when testing equality of variances (males higher).

(b)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.2855	0.5664	0.0426	2.4336	5.0000
M	359	4.1888	0.6756	0.0357	1.0000	5.0000
Diff (1-2)		0.0967	0.6417	0.0589		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	1.64	0.1014
Satterthwaite	Unequal	410.29	1.74	0.0824

P-value = 0.0082 when testing equality of variances (males higher).

(c)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.2515	0.6058	0.0455	2.1947	5.0000
M	359	4.1098	0.7088	0.0374	1.0000	5.0000
Diff (1-2)		0.1417	0.6765	0.0621		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	2.28	0.0230
Satterthwaite	Unequal	403.41	2.40	0.0166

P-value = 0.0186 when testing equality of variances (males higher).

(d)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.0381	0.5628	0.0423	2.3333	5.0000
M	359	4.0758	0.5859	0.0309	1.0000	5.0000
Diff (1-2)		−0.0377	0.5784	0.0531		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	−0.71	0.4783
Satterthwaite	Unequal	363.33	−0.72	0.4725

P-value = 0.5476 when testing equality of variances.

(e)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.1894	0.6386	0.0480	2.0000	5.0000
M	359	4.3100	0.5427	0.0286	2.0000	5.0000
Diff (1-2)		−0.1206	0.5761	0.0529		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	−2.28	0.0230
Satterthwaite	Unequal	304.65	−2.16	0.0317

P-value = 0.0104 when testing equality of variances (females higher).

(f)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.0063	0.5612	0.0422	2.0000	5.0000
M	359	3.9361	0.5868	0.0310	1.0000	5.0000
Diff (1-2)		0.0702	0.5785	0.0531		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	1.32	0.1869
Satterthwaite	Unequal	364.76	1.34	0.1805

P-value = 0.5042 when testing equality of variances.

(g)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.3116	0.4841	0.0364	2.3333	5.0000
M	359	4.2158	0.6059	0.0320	1.0000	5.0000
Diff (1-2)		0.0958	0.5687	0.0522		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	1.83	0.0672
Satterthwaite	Unequal	427.52	1.98	0.0486

P-value = 0.00084 when testing equality of variances (males higher).

(h)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.2013	0.4823	0.0362	3.0000	5.0000
M	359	4.1366	0.6250	0.0330	1.0000	5.0000
Diff (1-2)		0.0647	0.5819	0.0534		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	1.21	0.2265
Satterthwaite	Unequal	439.9	1.32	0.1874

P-value = 0.0002 when testing equality of variances (males higher).

(i)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.1465	0.5244	0.0394	2.6991	5.0000
M	359	4.0847	0.6461	0.0341	1.0000	5.0000
Diff (1-2)		0.0618	0.6086	0.0559		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	1.11	0.2694
Satterthwaite	Unequal	421.84	1.19	0.2363

P-value = 0.0018 when testing equality of variances (males higher).

(j)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.2926	0.4657	0.0350	2.6667	5.0000
M	359	4.3426	0.5114	0.0270	2.0000	5.0000
Diff (1-2)		−0.0500	0.4968	0.0456		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	−1.10	0.2736
Satterthwaite	Unequal	381.25	−1.13	0.2587

P-value = 0.1588 when testing equality of variances.

(k)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.0409	0.5240	0.0394	2.6250	5.0000
M	359	4.0269	0.6053	0.0319	1.0000	5.0000
Diff (1-2)		0.0141	0.5798	0.0533		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	0.26	0.7918
Satterthwaite	Unequal	398.9	0.28	0.7816

P-value = 0.0304 when testing equality of variances (males higher).

(l)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.2335	0.4775	0.0359	3.0000	5.0000
M	359	4.2077	0.5880	0.0310	1.0000	5.0000
Diff (1-2)		0.0259	0.5540	0.0509		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	0.51	0.6115
Satterthwaite	Unequal	421.68	0.55	0.5860

P-value = 0.0019 when testing equality of variances (males higher).

(m)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.2718	0.4791	0.0360	2.0000	5.0000
M	359	4.2398	0.5349	0.0282	1.0000	5.0000
Diff (1-2)		0.0320	0.5171	0.0475		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	0.67	0.5008
Satterthwaite	Unequal	386.95	0.70	0.4847

P-value = 0.0974 when testing equality of variances (males higher).

(n)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.2060	0.4996	0.0376	2.9554	5.0000
M	359	4.1444	0.5837	0.0308	1.0000	5.0000
Diff (1-2)		0.0615	0.5574	0.0512		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	1.20	0.2298
Satterthwaite	Unequal	402.88	1.27	0.2059

P-value = 0.0198 when testing equality of variances (males higher).

(o)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.1940	0.4350	0.0327	3.0000	5.0000
M	359	4.1558	0.5303	0.0280	2.0000	5.0000
Diff (1-2)		0.0382	0.5009	0.0460		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	0.83	0.4071
Satterthwaite	Unequal	418.01	0.89	0.3757

P-value = 0.0032 when testing equality of variances (males higher).

(p)

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	177	4.2277	0.4545	0.0342	2.6667	5.0000
M	359	4.1930	0.5268	0.0278	2.0000	5.0000
Diff (1-2)		0.0347	0.5041	0.0463		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	534	0.75	0.4545
Satterthwaite	Unequal	400.06	0.79	0.4319

P-value = 0.0268 when testing equality of variances (males higher).

has shown that student expect female faculty to not be as harsh on grading, with students giving female faculty lower evaluations on grading if they are harsh [12] [13] [15] [17] [18]. For question 10, "I understood how my grades were assigned in this course", the sample mean of class average responses was slightly lower for female faculty, although not significant (p-value = 0.2587). The class average responses for questions 12 and 13 on availability and providing feedback in a timely manner were actually slightly higher for female faculty, but not significant (p-values = 0.6848 and 0.8016, respectively). Recall that research suggests that students expect female faculty to be available more often than male faculty [12] [13] [15] [17] [18]. Research has also shown that males are rated higher in promptness even when taking the same amount of time to return assignments as female faculty [4].

We did consider the differences in variation of class average responses for each question between male and female instructors over the entire University (All Colleges). A significant difference in variability was found between class average responses for classes taught by male and female instructors for questions 5, and 15 at alpha equal to 0.05. A marginally significant difference was found between the class average responses for question 11 (p-value = 0.0906). The class average responses for question 5 had a larger variability for female faculty. This question was on rating the fairness of grading procedures. It is noted again that research has shown students expect female faculty to grade less harshly [12] [13] [15] [17] [18]. This could account for the larger variability among female faculty. The variances of the responses for male faculty were higher in the other two cases. The p-values for tests of variances are given in **Table 1**. The difference in variability of the class average responses for male and female faculty was tested was tested within the College of Science and Mathematics. A significant difference in variability between class average responses for classes taught by male and female instructors was found for questions 1, 2, 5, 7, 8, 9, 11, 12, 14, 15, and 16 at alpha equal to 0.05. A marginally significant difference in variability was found between class average responses for question 13 (p-value = 0.0974). In all cases, except for question 5, the class average responses were found to be less variable for classes taught by female instructors than for classes taught by male instructors (**Table 2**).

In the last phase of our research, we considered only classes in which at least 5 male students and 5 female students responded. Other research has found that male students tend to rate male instructors higher and female students tend to rate female instructors higher [5]-[7]. We wanted to test whether the proportion of classes taught by female instructors in which the male student response was higher was significantly lower than the proportion of classes taught by male instructors in which the male student response was higher. The sample proportion of classes taught by female instructors in which the male response was higher was calculated (Proportion 1) and the sample proportion of classes taught by male instructors in which the male response was higher (Proportion 2) was calculated. These are given in **Table 3**. In all cases, the sample proportion for female instructors was lower

Table 3. Proportion of classes in which male student response is higher.

Question	Proportion 1 (female)	Proportion 2 (male)	Test Statistic	P-value
1	0.4375	0.5432	−1.73	0.04182[*]
2	0.4821	0.5432	−1.25	0.10565
3	0.5089	0.5926	−1.37	0.08534
4	0.3750	0.5432	−2.79	0.00264[**]
5	0.4286	0.5617	−2.19	0.01426[**]
6	0.4821	0.6049	−2.02	0.02169[**]
7	0.4123	0.5185	−1.75	0.04006[*]
8	0.3947	0.5000	−1.74	0.04093[*]
9	0.4474	0.5802	−2.19	0.01426[**]
10	0.4561	0.5370	−1.27	0.10204
11	0.4298	0.5864	−2.59	0.00480[**]
12	0.4123	0.5000	−1.45	0.07353
13	0.4649	0.4877	−0.37	0.05715
14	0.4825	0.5864	−1.71	0.04363[*]
15	0.4561	0.5123	−0.87	0.19215
16	0.4298	0.5370	−1.77	0.03836

than the sample proportion for male instructors. There were 10 questions in which the male student response was significantly lower for female instructors at alpha equal to 0.05. These were for questions 1, 4, 5, 6, 7, 8, 9, 11, 14, and 16. There were 2 additional questions in which the male response was significantly lower at alpha equal to 0.10 for female instructors. These were for questions 3 and 12 with the proportions not significantly different at alpha equal to 0.10 for question 10, but with a p-value of 0.102. This is similar to the research of Feldman [5] who did find that there was a gendered interaction in student ratings as well as the research of Bachen *et al*. [11] who found that female students tended to rate female instructors higher. It is interesting to note that all of the sample proportion 1's are less than 0.50 except for the sample proportion association with question 3. Question 3 was the only question in which the sample proportion of classes taught by female instructors in which the male class average student response was higher, 0.5089, than the sample proportion of classes taught by female instructors in which the female class average student response was higher, 0.4911. Question 13 was the only question in which the sample proportion of classes taught by male instructors in which the male class average response was higher, 0.4877, was less than the sample proportion of classes taught by male instructors in which the female response was higher, 0.5123.

4. Conclusions

Our research did not find the gender indicator variable to be significant when proportion of students expecting A's and B's in model (All Colleges). We did find significant differences for questions 6, 11, and 15 with student responses associated with female faculty higher (understanding of course content, met or exceeded course objectives, instructor set and maintained high standards); this is when proportion of students expecting A's and B's is not controlled for in model. It does appear that a higher proportion of students taking courses from female faculty are expecting A's or B's. We also compared the variances of class average responses for male and female faculty across all colleges. The variances were significantly different or marginally significantly different for 3 of the 16 questions. The variance of class average responses was significantly higher for female faculty on question 5, the "fairness of procedures for grading". In the other two questions, the variance of class averages for male faculty was higher.

When considering only the College of Science and Mathematics, we did not find that female faculty was rated lower on average. We found female faculty to be associated with significantly higher ratings in creating an atmosphere conducive to learning (other research has found this), and communication ability of the instructor as a teacher (marginally). We did find that female faculty was rated significantly lower on the fairness of grading procedures. We did find that the variances of the class average responses between male and female faculty were significant or marginally significantly different on 13 of the 16 questions. In all of these cases, except for question 5 asking about the fairness of grading procedures, the variances were higher for male faculty class averages.

References

[1]   Read, W., Rama, D.V., & Raghunandan K. (2001). "The Relationship Between Student Evaluations of Teaching and Faculty Evaluations", *Journal of Education for Business*, 76, 189-192.

[2]   Arbuckle, J., & Williams, B.D. (2003). "Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations", *Sex Roles*, 49, 507-516.

[3]   Brady, K.L., & Eisler, R.M. (1999). "Sex and Gender in the College Classroom: A Quantitative Analysis of Faculty-Student Interactions and Perceptions", *Journal of Educational Psychology,* 91, 127-145.

[4]   Marcotte, Amanda (2014). "Best Way for Professor to Get Good Student Evaluations? Be Male.", *XXfactor,* Dec. 9, 2014.

[5]   Feldman, Kenneth. 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.

[6]   Feldman, Kenneth. 1992. "College Students' Views of Male and Female College Teachers: Part I-Evidence from the Social Laboratory and Experiments." Research in Higher Education 33:317-73.

[7]   Lueck, T.L., Endres, K.L, & Caplan, R. E. (1993). "The Interaction Effects of Gender on Teaching Evaluations", *Journalism Education,* 48, 46-54.

[8]   Worthington, A. (2002). "The Impact of Student Perceptions and Characteristics on Teaching Evaluations: A Case Study in Finance Education", *Assessment & Evaluation in Higher Education, 27, 49-64.*

[9]   Basow, Susan A. (1995). "Student Evaluations of College Professors: When Gender Matters", *Journal of Educa-*

tional Psychology, 87, 656-665.

[10] Centra, J.A. & Gaubatz, N.B. (2000). "Is There Gender Bias in Student Evaluations of Teaching?", *Journal of Higher Education,* 71, 17-33.

[11] Bachen, C.M., McLoughlin, M.M. and Garcia, S.S. (1999) Assessing the Role of Gender in College Students' Evaluations of Faculty. *Communication Education*, **48**, 193-210. http://dx.doi.org/10.1080/03634529909379169

[12] Basow, S.A. and Montgomery, S. (2005) Student Evaluations of Professors and Professor Self-Ratings: Gender and Divisional Patterns. *Journal of Personnel Evaluation in Education*, **18**, 91-106. http://dx.doi.org/10.1007/s11092-006-9001-8

[13] 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. http://dx.doi.org/10.1037/0022-0663.74.2.170

[14] Wilson, J.H., Beyer, D. and Monteiro, H. (2014) Professor Age Affects Student Ratings: Halo Effect for Younger Teachers. *College Teaching*, **62**, 20-24. http://dx.doi.org/10.1080/87567555.2013.825574

[15] Sprague, J. and Massoni, K. (2005) Student Evaluations and Gendered Expectations: What We Can't Count Can Hurt Us. *Sex Roles*, **53**, 779-793. http://dx.doi.org/10.1007/s11199-005-8292-4

[16] Laube, H., Massoni, K., Sprague, J. and Ferber, A.L. (2007) The Impact of Gender on the Evaluation of Teaching: What We Know and What We Can Do. *NWSA Journal*, **19**, 87-104.

[17] Sinclair, L. and 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. http://dx.doi.org/10.1177/0146167200263002

[18] Burns-Glover, A. and Veith, D. (1995) Revising Gender and Teaching Evaluations: Sex Still Makes a Difference. *Journal of Social Behavior and Personality*, **10**, 69-80.

[19] Jaschik, S. (2015) Rate My Word Choice. https://www.insidehighered.com/news/2015/02/09/new-analysis-rate-my-professors-finds-patterns-words-used-describe-men-and-women

[20] Felton, J. (2008) Attractiveness, Easiness and Other Issues: Student Evaluations of Professors on Ratemyprofessors.com. *Assessment and Evaluation in Higher Education*, **33**, 151-211. http://dx.doi.org/10.1080/02602930601122803

[21] Greenwald, A.G. and Gillmore, G.M. (1997) Grading Leniency Is a Removable Contaminant of Student Ratings. *American Psychologist*, **52**, 1209-1217. http://dx.doi.org/10.1037/0003-066X.52.11.1209

[22] Millea, M. and Grimes, P.W. (2002) Grade Expectations and Student Evaluation of Teaching. *College Student Journal*, **36**, 582-590.

[23] Culver, S. (2010) Course Grades, Quality of Student Engagement, and Students' Evaluation of Instructor. *International Journal of Teaching and Learning in Higher Education*, **22**, 331-336.

[24] Eizler, C.F. (2002) College Students' Evaluation of Teaching and Grade Inflation. *Research in Higher Education*, **43**, 483-501. http://dx.doi.org/10.1023/A:1015579817194

[25] Basow, S.A. and Martin, J.L. (2012) Bias in Student Evaluations. In: Kite, M.E., Ed., *Effective Evaluation of Teaching: A Guide for Faculty and Administrators*, Society for the Teaching of Psychology, Washington DC, 40-49.