Gender Bias in Professor Evaluations:
A Means Comparison analysis of Faculty Gender and Assignments: were integral in synthesizing course learnings

Paulina Velasquez
Educational Research Methods, Georgetown University
Dr. Doireann Renzi
27 April 2024

Research Question

This study investigates whether the gender of instructors has any relation with students' perceptions of the importance of assignments in synthesizing learning. By exploring potential variations in students' views based on instructor gender, the analysis aims to uncover biases or preferences that may impact attitudes toward gender of instructors and the assignments' importance in a class. A correlations analysis and means comparison will be used to assess correlations between instructor gender and student perception of assignments, contributing insights into pedagogical dynamics influenced by gender.

Hypothesis

Based on the research question, I assume that the results would indicate that there will be a statistically significant relationship between a faculty member's gender and whether students indicated that assignments were integral to their learning. Looking at the assumption deeper, I am hypothesizing that there might be a correlation between faculty members that identify as women and students that are more critical on their assignments and their instruction. Additionally, when exploring the results of the means comparison, I'm assuming there might be a notable difference between the average score between female and male instructors. This assumption comes from the literature review, which is featured below. Particularly the Laube article, discusses different studies that have found different gender based biases in student evaluations. And reflecting on my own experiences as a student and working in education, I assume that some students might be more critical of a female instructor than they are with male instructors.

Literature Review

My purpose in examining faculty gender and evaluation questions came from a suspicion that the data would present some bias in its results. My assumption is that female instructors would possibly face more scrutiny than their male counterparts. In my literature review I found several articles that gave me more depth and understanding of that assumption and helped shape my analysis of the results. Specifically, my literature review will look at gender-based biases, how gender can impact the classroom environment, student achievement, and discussion on intersectionality.

Gender-Based Bias in Literature

One of the primary sources I looked at was Heather Laube's Article, "The Impact of Gender on the Evaluation of Teaching: What We Know and What We Can Do." This article looked at several studies that centered around Teaching Evaluations and Gender on College Campuses. One study that they explored came to a conclusion that the complexities of gender expression and expectation play a large role in shaping biases. It states, "Some researchers argue it is not the sex of the professor but instead the degree to which the professor's personality meets or escapes traditional notions of gender that makes a difference in the kinds of ratings students give." (Luabe, Massoni, et. al, p. 89). In my discussion, I suggest that it is important to investigate further how these aspects might impact biases on evaluations. It also brings up the intricacies of gender expectations beyond identity. Instructors might identify as a certain gender but be exhibiting behavior or appearance that is outside a person's gender-expectations. This could easily impact someone's perception of a professor, thus impacting their evaluation. Further on this point, This article also mentions a survey conducted by Sprague and Massoni where students evaluated their professors by picking out four words to describe them. The study explains that students, especially in evaluating teachers they didn't like, use charged and emotional language (Luabe, Massoni, et. al, p.94). Expanding on this, in discussion with this survey, it concludes that female professors were often judged by their personality or by gender expectations while men were judged by their teaching performance. "Messner notes, "Students tend to judge their 'gender performance' and second by their teaching performance." (Luabe, Massoni, et. al, p.95). Putting both these aspects into consideration, it is clear that evaluations need to choose language that promotes thoughtfulness and avoids provoking or emotional language. I feel that the play data evaluation, while flawed in many ways, achieves this. The language is very neutral and isn't provoking or impassioned.

Exploring Intersectionality

Additional reading and studies examined biases on an intersectional basis. Beyond discussions of gender identity and expression, aspects like race and ethnicity may play big factors to students' biases. Both Dana Williams', "Examining the Relation between Race and Student Evaluations of Faculty Members: A Literature Review," and Benjamin Artz's "The Effect of Peer and Professor Gender on College Student Performance" seek to discuss these matters. Investigations or studies that keep intersectional values in mind allows for more nuance

and understanding. How white or caucasian female instructors are perceived can be very different from black female instructors. And those perceptions can be different if those instructors present queer or traditionally feminine. The Williams' article talks about a study that explores whether an instructor being a minority influences a student's evaluation. While the study didn't find evidence to support this main hypothesis they did learn that "the findings did suggest that an instructor's characteristics influence students' evaluations of which course material is considered controversial"(Williams, p.170). Perceiving something as controversial could just signify a discomfort and this could be based around instructor identity.

Personality and characteristics of course can be influenced by, and this be directly related with, who is rating them. The Artz article explores this concept. It studies student achievement and its relationship between same-gender or different-gender instructor student pairings. What was most significant about this study was in exploring that research question, it was able to track the gender of instructors that students sought for their courses and how the perception of their past courses and past instructors are influenced by that (Artz, Welsch p. 827). This point get's reiterated in an article by Holly Tatum. She states, "Interestingly, in that same study, there was a significant interaction between gender of professor and gender of student. Male students reported that they learned more from male professors while female students reported that they learned more from female professors." (Tatum, Schwartz, et. al p.749). Both these articles note the relationship between students' interaction with their courses and their instructors and the impacts of gender. It's no small leap to assume that both their preference for instructors and their experience in the classroom impact their evaluations of the class once they finish.

Conclusion

Ultimately, in doing the literature review there were many factors to consider in completing and reflecting on my analysis of the play data. One, I recognize the shortcomings of intersectional analysis considering I'm analyzing solely on reported gender on a binary. And in many cases, including the literature I read, many reported non-conclusive results in regards to biases solely based on gender. Even the Laube article brings in an assertion about these studies, "that direct effects of gender on evaluation were minor, trivial in size, and in the case of evaluations of actual professors, favored women rather than men." And then continues, "We argue that a more careful reading of the research literature reveals that the evidence is more mixed" (Luabe, Massoni, et. al, pp. 88-89). But I believe it is important to be open to the results

of this analysis. The research question and its variables are simple and might offer an answer that doesn't meet the complexities of the nuance of a broader issue. But I believe it is still worth exploring.

Details of the Variables in Question

The two variables I am exploring for this analysis are "Faculty Gender" and "Assignments: were integral in synthesizing course learnings." I ran a frequency analysis for both variables which are represented in *fig 1* and *fig 2*. The results of this analysis show the number of total respondents and a breakdown of the number of respondent for each value.

Figure 1 ***(Fig 1)***

Frequency Analysis of Variable "Faculty Genders"

Faculty Genders

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	915	49.4	54.1	54.1
	Male	768	41.4	45.4	99.5
	Not Reported	9	.5	.5	100.0
	Total	1692	91.3	100.0	
Missing	System	162	8.7		
Total		1854	100.0		

In looking at the variable data for faculty gender which is displayed in *fig 1*, there were altogether 1692 respondents (N=1692) to the survey. Of those respondents, 915 were reported as female, which is 54.1% of the data. Responses that are "Female" are valued as 1. Further, 768 respondents reported male which is 45.4% of the data. Responses that are reported as "Male" are valued as 2. Lastly, 9 respondents were "Not Reported" which was .5% of the data. "Not Reported" answers were re-coded out of the data, so the total Number of respondents was 1683 (N=1683). This was due to the specificity of the answer and the small number of respondents. The reasoning is explained further under "Changing Variables" Section.

Figure 2 ***(Fig 2)***

Frequency Analysis of Variable "Assignments: were integral in synthesizing course learnings"

Assignments: Were integral in synthesizing course learnings

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Almost Never	17	.9	1.1	1.1
	Seldom	40	2.2	2.6	3.7
	Sometimes	139	7.5	9.0	12.7
	Usually	355	19.1	23.0	35.6
	Almost Always	995	53.7	64.4	100.0
	Total	1546	83.4	100.0	
Missing	System	308	16.6		
Total		1854	100.0		

Now exploring the details of "Assignments: were integral in synthesizing course learnings," which is represented in *fig 2*, there were 1546 respondents in total (N=1546). Though after the exclusion of value three, "Sometimes," there are only 1407 responses (N=1407). But for the purposes of this section it will break down different variables using the original N=1546. The possible responses are "Almost Never," "Seldom," "Sometimes," "Usually," and "Almost Always." Starting in that order, 17 respondents reported "Almost Never" which was 1.1% of respondents. "Almost Never" is valued at 1. Next, 40 respondents reported "Seldom" which was 2.6% of the total and "Seldom" value was 2. 139 respondents reported "Sometimes" which is 9% of the total. These respondents were re-coded out of the data by listing it as "System-Missing." The justification for this is listed in the "Changing Variables" section. Further, "Usually" had 355 responses which was 23% of the total. The "Usually" response is valued at 4. Lastly, "Almost Always" was reported by 995 respondents which is 64.4% of the total. "Almost Always" is valued as 5.

Recoding Variables

In order to have a clearer understanding on the relationship and impact of the two variables, I re-coded the values for both.

First, I thought it was important to simplify the "Faculty Gender" data. I transformed this values to represent a binary; (1) for females and (2) for males. I removed responses for "Not Reported" from the data set. This was on account that out of 1692 responses only 9 responses answered with "Not Reported." With that small of a respondent pool, the results would have not been conclusive in analyzing that value and it could skew other data. Additionally, "Not

reported" doesn't hold much significance in meaning and I wouldn't know how to properly analyze it. I believe it is important to make a quick note that this data set has excluded people who identify outside the gender binary and this is the reason that this report is looking at gender specifically in a binary.

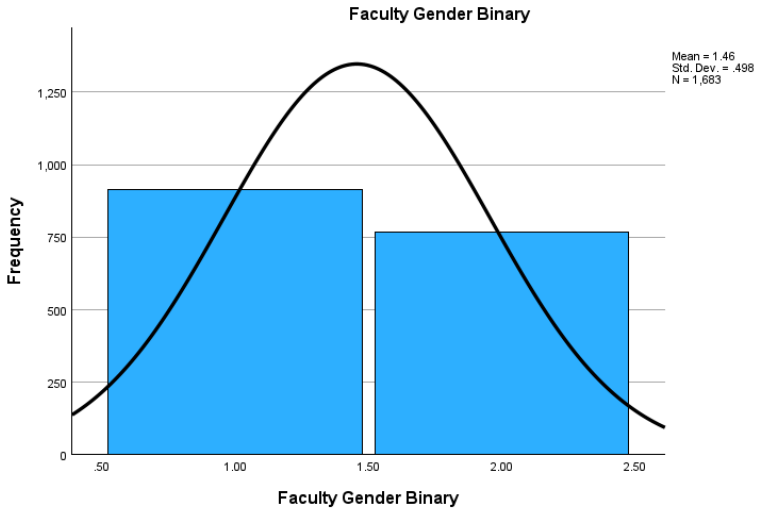
In regards to changing the values of the second variable, I decided that it was best to eliminate the "sometimes" option response from the data set. Altogether,139 responses were re-coded out of the data. Ultimately, for analyzing this data, I wanted to understand and break down students' ratings for "Assignments were integral to synthesizing student learning" as responses that were favorable or unfavorable. Because the option "Sometimes" is a neutral answer, it did not offer insight to that extent so it was removed.

Descriptives of the Variables of Interest

The following data and their accompanying figures depict a Descriptives Analysis that I ran to look at the Mean and Standard Deviation for each variable.

Figure 3 (*fig 3*)

Descriptive Analysis of "Faculty Gender" represented by a Histogram.

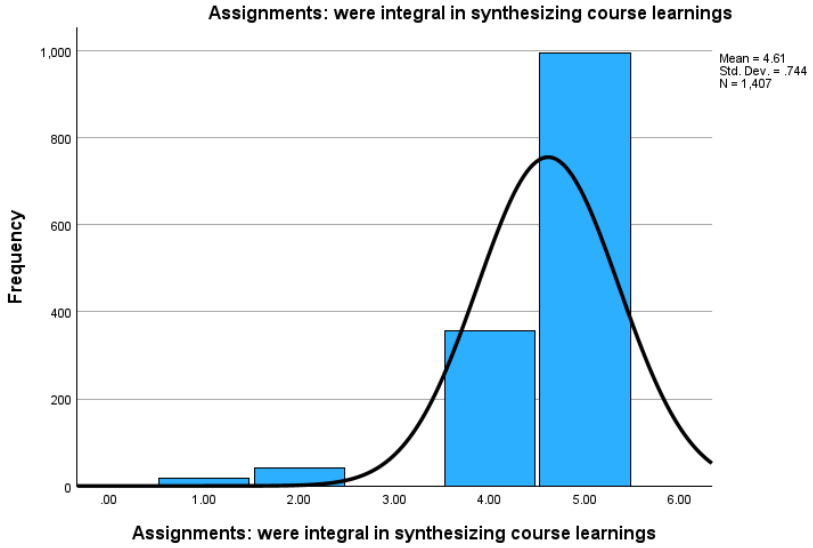


Looking at *Fig 3*, the histogram shows the results of the descriptive analysis and the results are as follows, (N=1683), (M=1.46), and (SD=0.498). From the 1683 faculty members the average response was 1.46 . Since I've coded females as 1 and males as 2, the mean of 1.46 states that, on average, the faculty population of the dataset has more females than males. This

can be ascertained because the average is slightly closer to 1 (1=female) than to 2 (2=male). This just reiterates what was seen on the frequency analysis. The standard deviation of the Faculty Gender variable is 0.498 . A low standard deviation 0.498 means that reported gender in the dataset are relatively close to the mean value of 1.46. This makes sense because in this data set gender is being studied on a binary.

Figure 4 (*fig 4*)

Descriptive Analysis of "Assignments: were integral in synthesizing course learnings" represented by a Histogram.



Looking at *fig 4*, the data shows the mean value is 4.61 (m=4.61). This indicates that the average response to this statement across all respondents was 4.61. A mean of 4.61 suggests that, on average, respondents agreed that assignments were "usually" or "almost always" integral in synthesizing the course learnings. Further, the standard deviation is 0.744 (SD=0.744). A standard deviation of 0.744 indicates that most responses were clustered relatively close to the mean of 4.61, but there was still some variation in how respondents rated the importance of assignments. To summarize, the majority of respondents viewed their assignments as an important aspect of consolidating what they learned in the course.

Analysis Justification

In order to analyze my research question, the best tests to use are a correlation analysis and a means comparison. The correlation analysis allows me to see if there is any statistically

significant relationship between the two variables. I will use this as a baseline to see if there is anything to explore further. What I will find most interesting are the results from the Means Comparison done by a Means Analysis. That way I can dissect the differences between female and male instructors and their performance in the specific evaluation question. I will be able to see what the average score is for both and see if there is any notable difference.

Analysis

Figure 5 (*fig 5*).

Results Table of a Correlation Analysis between "Faculty Gender" and "Assignments: were integral to synthesizing Course Learning"

Correlations

		Faculty Gender Binary	Assignments: were integral in synthesizing course learnings
Pearson Correlation	Faculty Gender Binary	1.000	-.088
	Assignments: were integral in synthesizing course learnings	-.088	1.000
Sig. (1-tailed)	Faculty Gender Binary	.	<.001
	Assignments: were integral in synthesizing course learnings	.001	.
N	Faculty Gender Binary	1399	1399
	Assignments: were integral in synthesizing course learnings	1399	1399

In exploring the correlations between these two variables, I ran a correlations analysis which is represented in *fig. 5.* The results are as follows (N=1399), ($p<.001$), ($r = -0.088$). To break this down, the sample size of respondents that answered both questions, "Faculty Gender" and "Assignments were integral in synthesizing course learnings" was 1399. Next, looking at the statistical significance, the p-value which equals .001 indicates that the data is statistically significant, and the relationship is unlikely to be random. Exploring the actual correlation coefficient, it equals -0.088** ($r = -0.088$). This was a very interesting finding because the correlation coefficient is negative it shows that there is a negative correlation between Faculty Gender and whether Assignments: were integral in synthesizing course learnings. In other words, this analysis shows that as the perception of assignments' importance decreases, the proportion of males (coded as 2) in the sample tends to slightly increases, and vice versa. To be clear, a correlation of -0.088 is fairly minimal but is still statistically significant.

Figure 6 (*fig 6*).

Results Table of a Means Analysis between "Faculty Gender" and "Assignments: were integral to synthesizing Course Learning"

Report

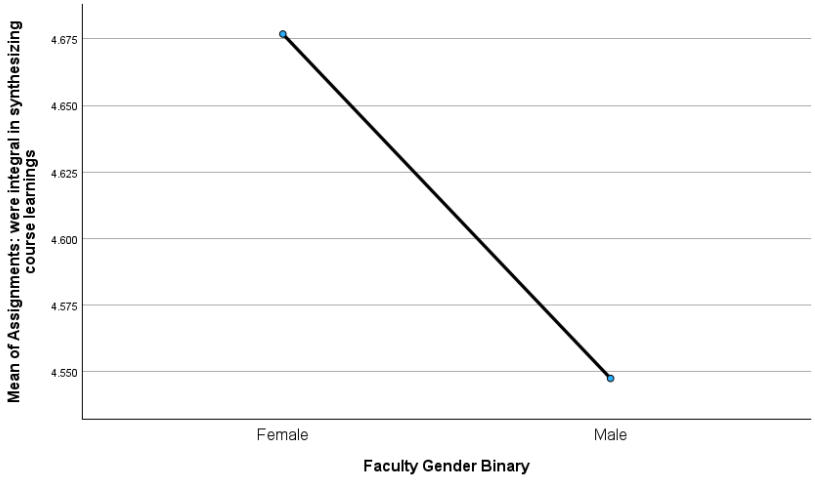
Assignments: were integral in synthesizing course learnings

Faculty Gender Binary	Mean	N	Std. Deviation
Female	4.6768	789	.68020
Male	4.5475	610	.78901
Total	4.6204	1399	.73218

To explore this relationship further, I ran a Means Comparison via a Means Analysis. I wanted to see what specifically the correlation is showing. The results of this test are depicted in *fig 6* & *fig 7,* depicted below. In *fig 6*, the chart shows the test results for Female Instructors (N=789), (M=4.677), and (SD=.680). So in total there were 789 female instructors featured in this test and the average answer was 4.677 which shows a very high average rating for assignments that were integral for course learnings. Switching over to the results for Male Instructors, the chart reads (N=610), (M=4.55), and (SD=.789). So for the 610 Male Instructors, they received on average a score of 4.55 which is also very high. In comparison between those two values, Female and Male professors are performing very similarly on this question on the survey. Focusing in on *fig 7*, this line graph depicts this difference. Ultimately, the difference between the means is only 0.127.

Figure 7 (*fig 7*).

Line graph of the Results of a Means Analysis between "Faculty Gender" and "Assignments: were integral to synthesizing Course Learning"



Paying attention to the variables on the y-axis on *fig 7*, the mean score for women is slightly above 4.675 line and the mean score for men is slightly below 4.55, which is why the difference appears dramatic. If this chart was blown-out and the Y-axis values ranged from 1 to 4, this difference would present fairly miniscule.

Results of Analysis

In completing this analysis, my original hypothesis was proven wrong. I initially thought that the results would show that responses to the survey would show a bias against female instructors. Looking at the question, "Assignments: were integral to course learning" the results of the correlation analysis showed the opposite. The correlation, ($r = -.088$), shows that there is a bias, though very slight, against male instructors. Again, when we examine that further, using a means comparison, it is very clear that the difference being spotted in correlation analysis is very slight. Again there is only .127 difference between the averages of the scores for that question. This could be very normal results for due variability of one question. If this same result was replicated after several tests then that would be something the university would want to look at. Altogether, both male and female instructors were rated very highly by students who answered that question.

Discussion

My interest in looking at these variables and asking this question was to explore possible sexism against female instructors in these evaluations. In regards to what was found in this analysis, though fairly minute, there is a slight bias against men. It would probably be pertinent to explore other questions and see if this is a pattern that continues. Or to repeatedly look at this question every semester and see if this is a consistent result. But exploring this topic further, there are many factors in regards to identity that can cause a bias against instructors. As Laube's article notes, instructors can face harsher scrutiny because of gender, race, ethnicity, language, etc (Laube, Massoni, et.al). Exploring where these biases might exist is important when considering how these evaluations might impact instructors. Also, this shows the importance of thoughtful formatting and question writing for evaluations. It is important to ensure that students or evaluators are more reflective of their answers, hopefully leading to more accurate data of instructor performance.

Resources

Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluations. Research in Higher Education, 58(4), 341–364. http://www.jstor.org/stable/26451440

Laube, H., Massoni, K., Sprague, J., & Ferber, A. L. (2007). The Impact of Gender on the Evaluation of Teaching: What We Know and What We Can Do. NWSA Journal, 19(3), 87–104. http://www.jstor.org/stable/40071230

Number and Growth of Faculty, by Gender and Rank: Selected Years (. (2005). https://www.acenet.edu/Documents/FactSheet-Number-and-Growth-of-Faculty-by-Gender-and-Rank-Selected-Years-2005-2007-2009.pdf

Tatum, H. E., Schwartz, B. M., Schimmoeller, P. A., & Perry, N. (2013). Classroom Participation and Student-Faculty Interactions: Does Gender Matter? The Journal of Higher Education, 84(6), 745–768. http://www.jstor.org/stable/43694532

Williams, D. A. (2007). Examining the Relation between Race and Student Evaluations of Faculty Members: A Literature Review. *Profession*, 168–173. http://www.jstor.org/stable/25595863