Juliet Bu
GOV 1372
Final Paper
December 12, 2025

Evaluating Gender Bias in Harvard Course Evaluations

Introduction

Recent reports depict women continuing to outpace men in college enrollment and graduation, and as more women have obtained a college education, women's representation in university faculty has increased as well (Nietzel, 2024; AAUP, 2020). However, these increases in the number of women faculty members, while a positive sign for gender equality in the higher education space, do not necessarily reveal the full nuance of women faculty members' experiences in universities. The academic literature suggests that students display gender bias when evaluating university faculty, shown by disparities in evaluation ratings for male versus female course faculty. The relationship between gender bias and political ideology is well-documented. Gender bias from university students therefore has a number of political implications. A recent working paper by Mathisen (2025) finds that the ideological gender gap in Norwegian high schoolers has surged over the past decade, with the teenage age group becoming the most gender-polarized demographic (p. 4). Mathisen argues that this polarization is driven by sexist attitudes of skepticism towards gender equality from young men (p. 22). Skepticism towards gender equality is becoming increasingly associated with right-wing politics, accounting for 40-50 percent of the increased polarization of the Norwegian high school students (p. 22).

By analyzing student course evaluations at Harvard, this paper examines how the micro foundations of the gender-ideological gap manifests itself in an everyday setting—a college campus. Harvard is a particularly unique institution to study, given its presence in the American political dialogue and tendency to produce future societal leaders. If there is evidence of gender bias among Harvard's student body, Mathisen's findings raise the question of whether this sexism could drive further political polarization or perpetuate gender bias in society at large, especially if students enter positions of power. By examining student course evaluations—where subconscious expressions of sexism, in particular, may appear—this paper considers the prevalence of gender bias at Harvard and what its implications are for political ideology.

The data in this study are from the most recently completed set of QGuide course evaluations for all spring 2025 courses. After cleaning, the data contains evaluation scores for 1,044 courses. From the analysis, I find no significant relationship between lecturer gender and lecturer score or course score. I also find there is no significant difference in the relationship between lecturer gender and lecturer score across department types.

Motivation

It is widely recognized in the literature that gender bias manifests itself in professor evaluations. Empirically, researchers have conducted many studies controlling for teaching quality, still finding bias. Mengel, Sauermann, and Zölitz (2017) use a natural experiment with random teacher assignment to reveal systematically lower evaluations for women, despite finding that grades and workload did not vary by instructor gender (p. 5–6). Mitchell and Martin (2018) also ran an empirical study, theirs focused on online coursework, reporting higher evaluation scores for the male instructor even for questions that were not clearly instructor-specific (p. 648-652). These empirical findings demonstrate causal evidence that women instructors receive lower ratings, a pattern that is not merely explained by confounding variables. Through a qualitative analysis, Mitchell and Martin's study also found that women instructors were more likely to be evaluated based on personality and appearance, and more likely to be referred to in lower-status terms like "teacher" rather than "professor" (p. 648-651), showing how gendered norms about women and men frame the aspects of an instructor's performance that the student evaluator focuses on. In considering the driving forces behind gender disparities, Mengel, Sauermann, and Zölitz find that the gender disparity is particularly driven by male students, is more pronounced in math courses, and is especially harsh for more junior female instructors (p. 1-2; p. 32-33). Owen, De Bruin, and Wu (2024) find that the evaluation of female instructors received a larger penalty for harsher grading as well, demonstrating the double gendered standard female instructors face. Finally, Saygin and Zhang (2025) find that even if responses to component evaluation questions are equal, female instructors may still receive a lower overall score, showing a "residual bias" manifesting itself when students are asked for a more general, less specific rating (p. 9). Saygin and Zhang further emphasize that overall ratings are most salient to prospective students and university decision-makers, suggesting that ratings could entrench institutional gender bias against female instructors progressing in academia.

This research paper extends the existing findings in the literature to assess their applicability to Harvard course evaluations and to consider the relationship between Harvard students' gender bias and political ideology. As it relates to political ideology, conservatism at Harvard in particular has faced backlash for their rejection of women (Gerstein and Reardon, 2025). This attitude suggests that gender bias and right-wing ideology may be intrinsically linked among the Harvard student body, similar to what Mathisen found in the study of Norwegian high school students. Harvard has also faced a shifting political climate. Despite having its first woman of color president in 2023, President Gay quickly lost her new leadership title mere months into the presidency, with some blaming sexism for her downfall (Harper, 2024). Additionally, throughout the past two years, the Harvard administration has faced heavy scrutiny and criticism from political figures, especially those in the Republican Party. These political dynamics at Harvard suggest that the Harvard student body is an interesting subject for a study of gender bias and political ideology.

Hypotheses

I hypothesize that on average, female faculty members are given lower QGuide lecturer and course scores than male faculty members, when controlling for potential confounders like course score, enjoyment, workload, and department. Additionally, the relationship between gender and evaluative scores will vary between department types. A discrepancy between how female and male faculty are evaluated is evidence of gender bias in the Harvard student body, and these gender bias attitudes may be correlated with a shift towards more right-wing political ideologies.

Data Description

The dataset used in this paper is the spring 2025 semester QGuide report for every undergraduate course at Harvard. The data are collected via end-of-semester online surveys that ask students a series of quantitative and qualitative evaluative questions about the course and teaching staff. This dataset is courtesy of Jay Chooi, a Harvard undergraduate, who scraped the QGuide for raw numerical scores and additionally used natural language processing to add variables not directly presented in the QGuide itself. The spring 2025 semester QGuide is the most recent set of student evaluations available, which I argue is therefore the most accurate representation of gender bias among the current student population. The sample is sufficiently large—at 1,044 observations after cleaning—suggesting that the analysis is not weakened by only focusing on one semester. Furthermore, given how volatile the political climate around Harvard was in spring of 2025 due to its conflict with the Trump Administration, it seemed appropriate to focus only on the spring 2025 semester, and not in addition to earlier semesters, to account for any potential confounding effects from the drastic shift in the political climate.

The original sample contains 1,567 courses, each with their own QGuide report. Each course includes identifying information—course code, title, teacher, a link to the QGuide, the FAS code, its unique code, and the course_id. To the data, I also added variables for faculty first name and gender. First names were scraped by a Python script that accessed the QGuide HTML from the links in the dataset and extracted the faculty first name. Gender was assigned with help from an LLM using a database of male and female names, though the faculty name and gender are not entirely accurate due to limitations in data processing and knowledge. The dataset also contains information about the number of student respondents and total students in the class, which I used to create a response rate variable that was used in data cleaning. Then, the mean, median, and standard deviation is reported for 6 evaluation variables: the course score, the lecturer score, the workload score (measured in hours of workload), the recommendation score, the sentiment score (coded by Chooi based on a natural language processing of QGuide comments), and the "gem probability" (also coded by Chooi based on the comments to represent the likelihood of a course being an "easy A"). Course score, lecturer score, workload score, and recommendation score all take on values between 1-5. For this study, I specifically focus on the mean of these scores, given that the median would always be a discrete number and thus not as informative for regression analysis. Sentiment score was measured on a scale of -1 to 1, though the values in this dataset remained within the -0.4 to 1 range, presumably because there were no universally negative course evaluations. The gem probability score is an evaluation of how likely a course is to be an "easy A," but, as will be discussed, largely takes on the value of zero and is therefore not considered in this data analysis. Finally, the dataset contains the best, worst, and most gem-classifying comments—reported as qualitative entries with the text of the comment in the dataset—in addition to the maximum and minimum sentiment score, all created by Chooi through natural language processing of the QGuide qualitative student comments.

Analysis:

To clean the data, I first dropped any observations where the values of my variables of interest were unknown or N/A: the instructor's gender, some of which were coded as unknown by the LLM because those names were not located in the LLM's database of names, the lecturer score (an outcome variable), the course score (another outcome variable), the sentiment score, the gem probability, the workload, and the recommendation score. I also dropped any observations where the response rate was less than 60% (in line with best statistical practice on survey response rate), to ensure the QGuide was an accurate representation of student sentiments (Fincham, 2008). An initial review of the distribution of outcome variables showed a substantially large number of entries with zero values for lecturer score. Upon further investigation of the raw data, it seems that when an instructor had multiple entries for the same course due to different sections of the course taught, all sections would be reported as a zero lecturer score. It was also impossible for lecturer mean score to take on a value of zero, as in the survey itself, respondents rated instructors by excellent, coded as a score of 5, very good, coded as a score of 4, good, coded as a score of 3, fair, coded as a score of 2, and unsatisfactory, coded as a score of 1, with no option for a student response to be coded as a 0. As such, I decided to additionally drop any observations where the mean lecturer score was 0. After dropping observations, I manually sorted each department into a department type, as reported by Harvard in the course catalog: Arts/Humanities, Social Sciences, Sciences/Engineering, Languages, and Miscellaneous courses, like General Education or Expository Writing courses. Using the course prefixes, I assigned each course a department type, used for further analysis of disparities between department types. The data cleaning reduced the number of observations in the dataset from 1,567 original observations to 1,044 cleaned observations.

My analysis began with exploratory data analysis through visualizations and summary statistics. I first examined the distribution of all of my outcome and control variables to understand the skew of the data and whether there were outliers (Fig. 1). I found that most of my variables were left skewed, except for the gem probability and workload. The left skew largely made intuitive sense, as the university would probably select out any courses that had received notably low scores from a previous semester, ensuring evaluation scores remained left skewed. Notably, gem probability distribution showed that an overwhelming majority of gem probability scores were zero, which led to the decision to exclude gem probability from the regression analysis.

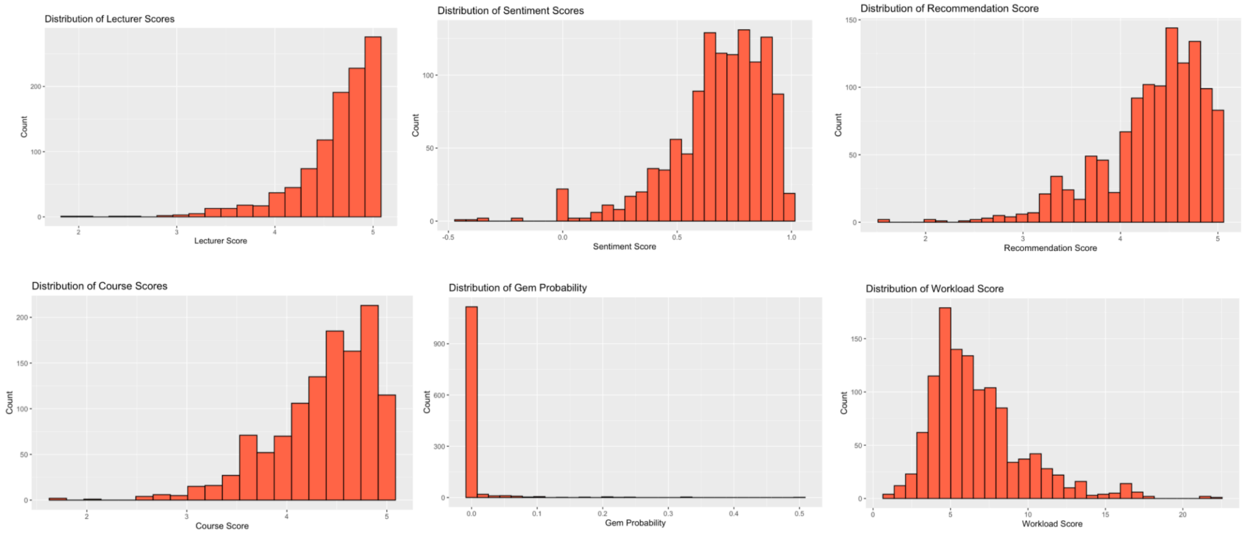


Figure 1: Histogram for Distribution of Outcome and Control Variables. Data Source: Harvard QGuide Spring 2025

I then examined the proportion of female and male instructors at the university level and by department type (Fig. 2). I found that at Harvard in general, lecturers are 37.2% female and 62.8% male. Examining gender proportions by department, Science/Engineering has the most disparity, with 30.9% female lecturers and 69.1% male lecturers. Languages have a disparity but in the opposite direction, with 62.7% female lecturers and 37.3% male lecturers. Finally, Arts/Humanities are the most equal in gender representation, with 45.3% female lecturers and 54.7% male lecturers.

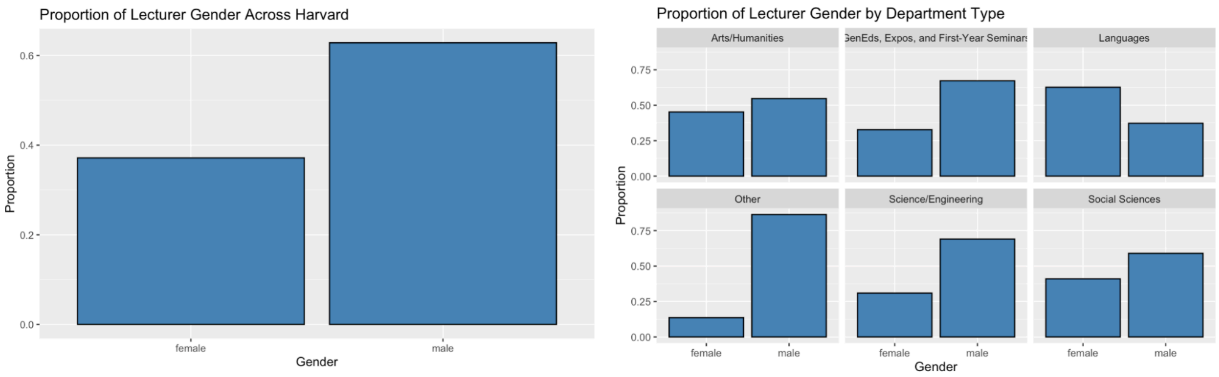


Figure 2: Bar plot of Proportions of Lecturer Gender for Harvard at Large and by Department. Data Source: Harvard QGuide Spring 2025

Next, I created two violin plots of my outcome variables by department type, to understand how the distribution of these outcome variables differed between department types (Fig. 3).

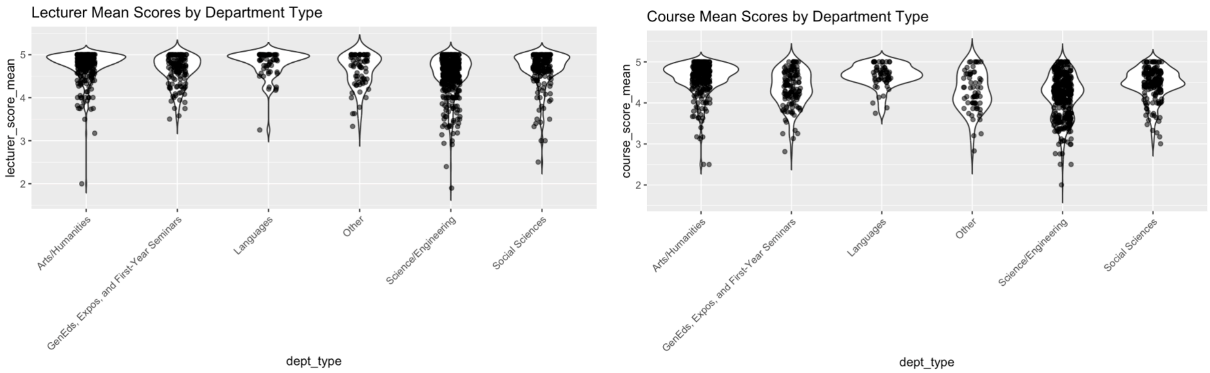


Figure 3: Violin Plots of Distribution of Lecturer Mean Score and Course Mean Score by Department Type. Data Source: Harvard QGuide Spring 2025

For both lecturer mean score and course mean score, Science/Engineering had the greatest distribution, suggesting more variability across courses in this department type. It is important to note that the Science/Engineering department category contains the largest number of classes, given this department type encompasses a vast array of courses, from classes for pre-med students to statistics courses. Another notable aspect of this visualization is that in both Arts/Humanities and Social Sciences, both lecturer score and course mean seem densely clustered at a high value, between 4-5, suggesting a greater consistency in experience in these department types and higher overall satisfaction in these department types compared to others.

The final exploratory visualization I created was a correlation matrix, to gain a preliminary understanding of the relationship between my outcome and control variables (Fig. 4). One takeaway that stood out from the correlation matrix was the strong correlation (0.93) between the recommendation score and the course score. Additionally examining how similarly these two variables were correlated with the other variables as well, the correlation matrix suggested potential collinearity between the course score and recommendation score variables. As a result, I decided to remove recommendation score from all regressions. For the regression of lecturer score on gender, I removed recommendation score as a control, because course score was already controlled for. For the regression of course score on gender, I removed recommendation score as a control to ensure the close correlation did not dilute the effect of my variable of interest on course score. Finally, in the interaction term regression examining the relationship between gender and lecturer score by department, I similarly removed recommendation score as a control, because course score was already controlled for.

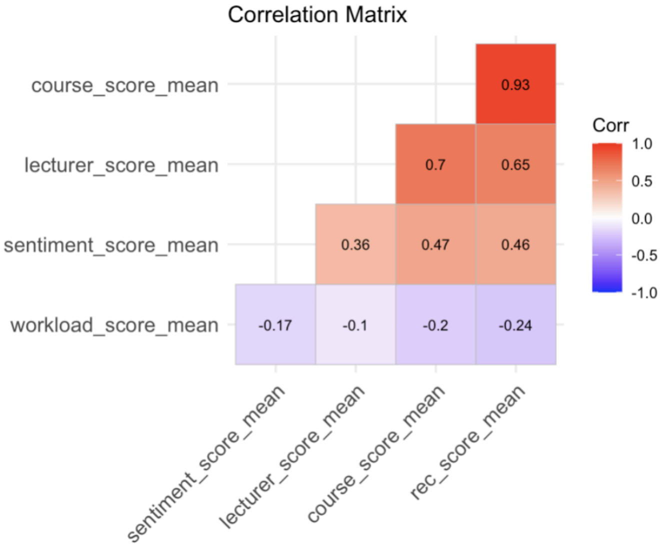


Figure 4: Correlation Matrix of Outcome and Control Variables. Data Source: Harvard QGuide Spring 2025

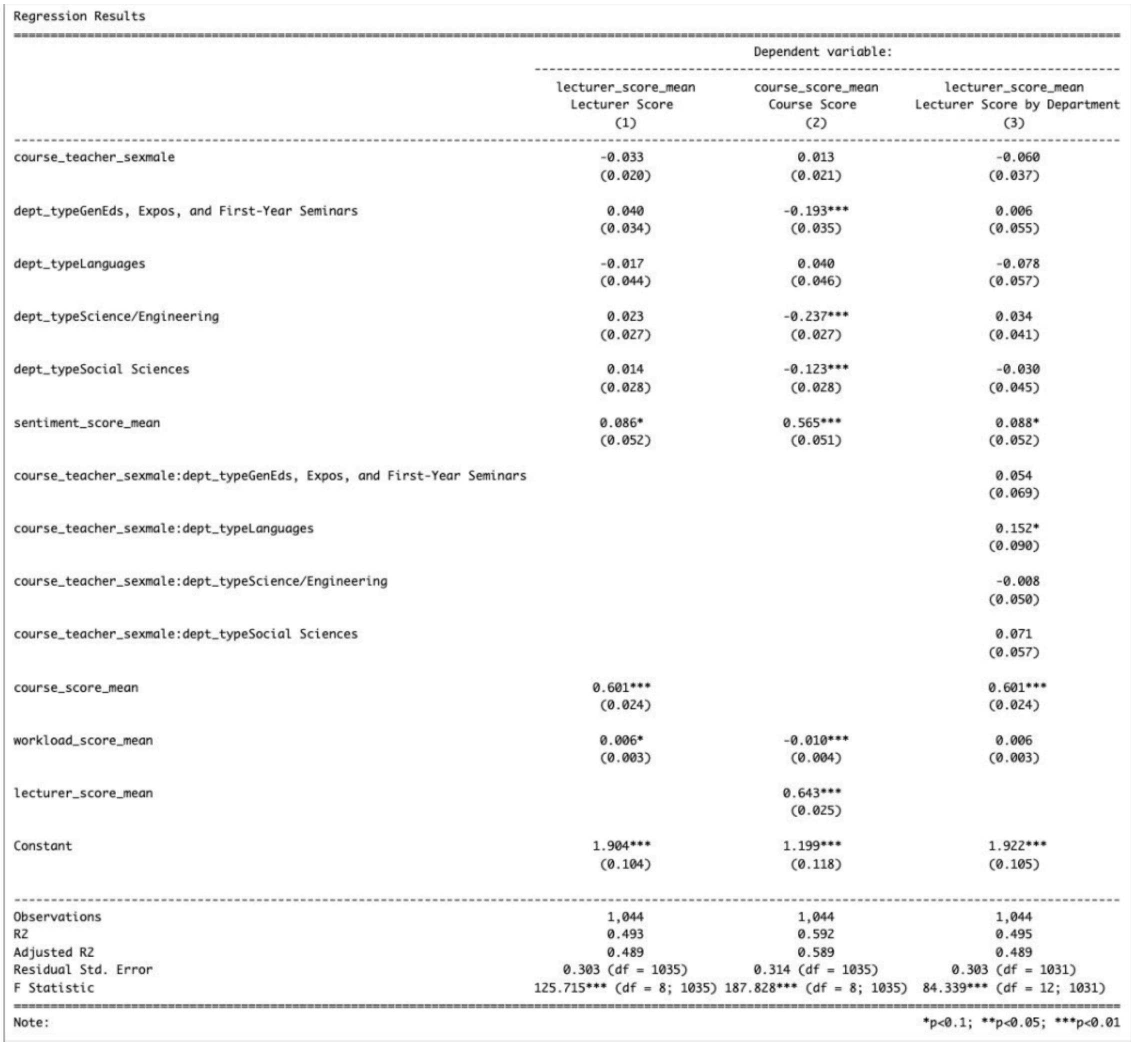
For my regression analysis, I ran three primary regressions to test my hypotheses (Fig. 5). The first was a linear regression of gender on the mean lecturer score, controlling for department, sentiment score, course score, and workload score—all available evaluation variables in the QGuide other than gem probability and recommendation score, which were taken out as a result of the exploratory phase. Because my variables are continuous on a scale of 1-5, the linear regression model was appropriate for evaluating the correlation between my predictor and outcome variable. The second regression was a linear regression of gender on the mean course score, controlling for department, sentiment score, course score, and workload score. I chose to consider course score as an outcome variable in addition to lecturer score to understand whether lecturer gender might have a significant effect on both the evaluation of the individual and the course. The third regression was a linear regression with an interaction term of gender and department type to understand whether the relationship between gender and lecturer score varied across departments. The third regression had the same controls as the first and second.

Results

The regressions did not yield any statistically significant relationship between professor gender and lecturer score or professor gender and course score (Fig. 5). The regression of lecturer score on gender suggests a potentially slightly negative relationship, where male lecturers receive, on average, lecturer scores that are 0.033 points lower than female lecturers when controlling for department type, sentiment score, course score, and workload score, though this coefficient is not statistically significant. The regression of course score on gender had a positive coefficient of 0.013, meaning that male lecturers receive, on average, course scores that are 0.013 points higher than female lecturers when controlling for department type, sentiment score, workload score, and lecturer score, though this effect is also not statistically significant (Fig, 5). The failure to find a statistically significant result in both the regression of lecturer score and course score on gender rejects my hypothesis that there is systematic gender bias in how Harvard students quantitatively evaluate lecturers and courses.

The regression of lecturer score on gender by department shows a weakly statistically significant coefficient (at the $p < 0.1$ level), 0.152, on the course_teacher_sexmale:dept_typeLanguages variable, meaning male professors in the Languages department are evaluated slightly more favorably than female instructions, on average, compared to the Arts and Humanities department (the baseline), when controlling for department type, sentiment score, course score, and workload score (Fig. 5). This result is interesting, because the Language department has the largest female to male instructor ratio, with

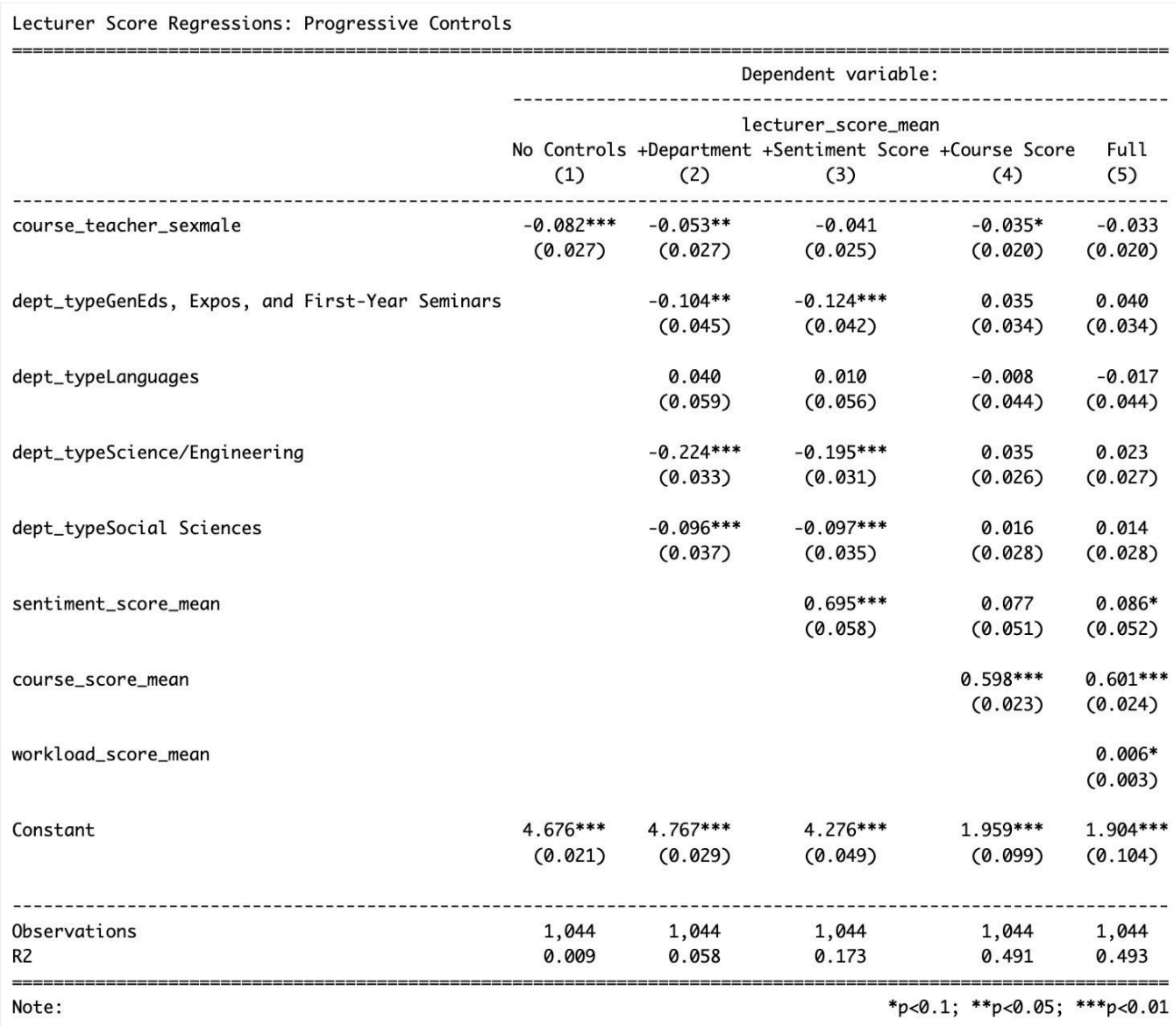
62.7% female lecturers and 37.3% male lecturers. This finding could suggest that where there are more female instructors, they are more likely to be reviewed slightly harsher than male counterparts, though in the context of the 5-point scale, the numerical difference is not vastly meaningful. Alternatively, the finding from the first regression that male instructors receive, on average, slightly lower lecturer scores may suggest that because there are more male lecturers at Harvard, they are probabilistically likely to receive more negative ratings than female counterparts. A similar probabilistic interpretation could explain why, because the Languages department has more female lecturers than male lecturers, there seems to be a greater gender disparity in lecturer scores in favor of male lecturers in the Languages department. Other than the Language department, however, these findings reject my hypothesis: I do not find systematic differences in how lecturer gender impacts student-given lecturer scores across departments.



```
Regression Results
==================================================================================================================================================
                                                                                                  Dependent variable:
                                                                        --------------------------------------------------------------------------
                                                                        lecturer_score_mean        course_score_mean          lecturer_score_mean
                                                                          Lecturer Score             Course Score         Lecturer Score by Department
                                                                               (1)                        (2)                        (3)
--------------------------------------------------------------------------------------------------------------------------------------------------
course_teacher_sexmale                                                       -0.033                     0.013                      -0.060
                                                                            (0.020)                    (0.021)                    (0.037)

dept_typeGenEds, Expos, and First-Year Seminars                              0.040                    -0.193***                    0.006
                                                                            (0.034)                    (0.035)                    (0.055)

dept_typeLanguages                                                           -0.017                     0.040                      -0.078
                                                                            (0.044)                    (0.046)                    (0.057)

dept_typeScience/Engineering                                                 0.023                    -0.237***                    0.034
                                                                            (0.027)                    (0.027)                    (0.041)

dept_typeSocial Sciences                                                     0.014                    -0.123***                    -0.030
                                                                            (0.028)                    (0.028)                    (0.045)

sentiment_score_mean                                                         0.086*                    0.565***                    0.088*
                                                                            (0.052)                    (0.051)                    (0.052)

course_teacher_sexmale:dept_typeGenEds, Expos, and First-Year Seminars                                                             0.054
                                                                                                                                  (0.069)

course_teacher_sexmale:dept_typeLanguages                                                                                          0.152*
                                                                                                                                  (0.090)

course_teacher_sexmale:dept_typeScience/Engineering                                                                                -0.008
                                                                                                                                  (0.050)

course_teacher_sexmale:dept_typeSocial Sciences                                                                                    0.071
                                                                                                                                  (0.057)

course_score_mean                                                           0.601***                                              0.601***
                                                                            (0.024)                                               (0.024)

workload_score_mean                                                          0.006*                   -0.010***                    0.006
                                                                            (0.003)                    (0.004)                    (0.003)

lecturer_score_mean                                                                                    0.643***
                                                                                                       (0.025)

Constant                                                                    1.904***                   1.199***                   1.922***
                                                                            (0.104)                    (0.118)                    (0.105)

--------------------------------------------------------------------------------------------------------------------------------------------------
Observations                                                                 1,044                      1,044                      1,044
R2                                                                           0.493                      0.592                      0.495
Adjusted R2                                                                  0.489                      0.589                      0.489
Residual Std. Error                                                    0.303 (df = 1035)          0.314 (df = 1035)          0.303 (df = 1031)
F Statistic                                                       125.715*** (df = 8; 1035) 187.828*** (df = 8; 1035) 84.339*** (df = 12; 1031)
==================================================================================================================================================
Note:                                                                                                                  *p<0.1; **p<0.05; ***p<0.01
```

Figure 5: Regression Table of Mean Lecturer Score, Mean Course Score, and Mean Lecturer Score by Department, on Lecturer Gender. Data Source: Harvard QGuide Spring 2025

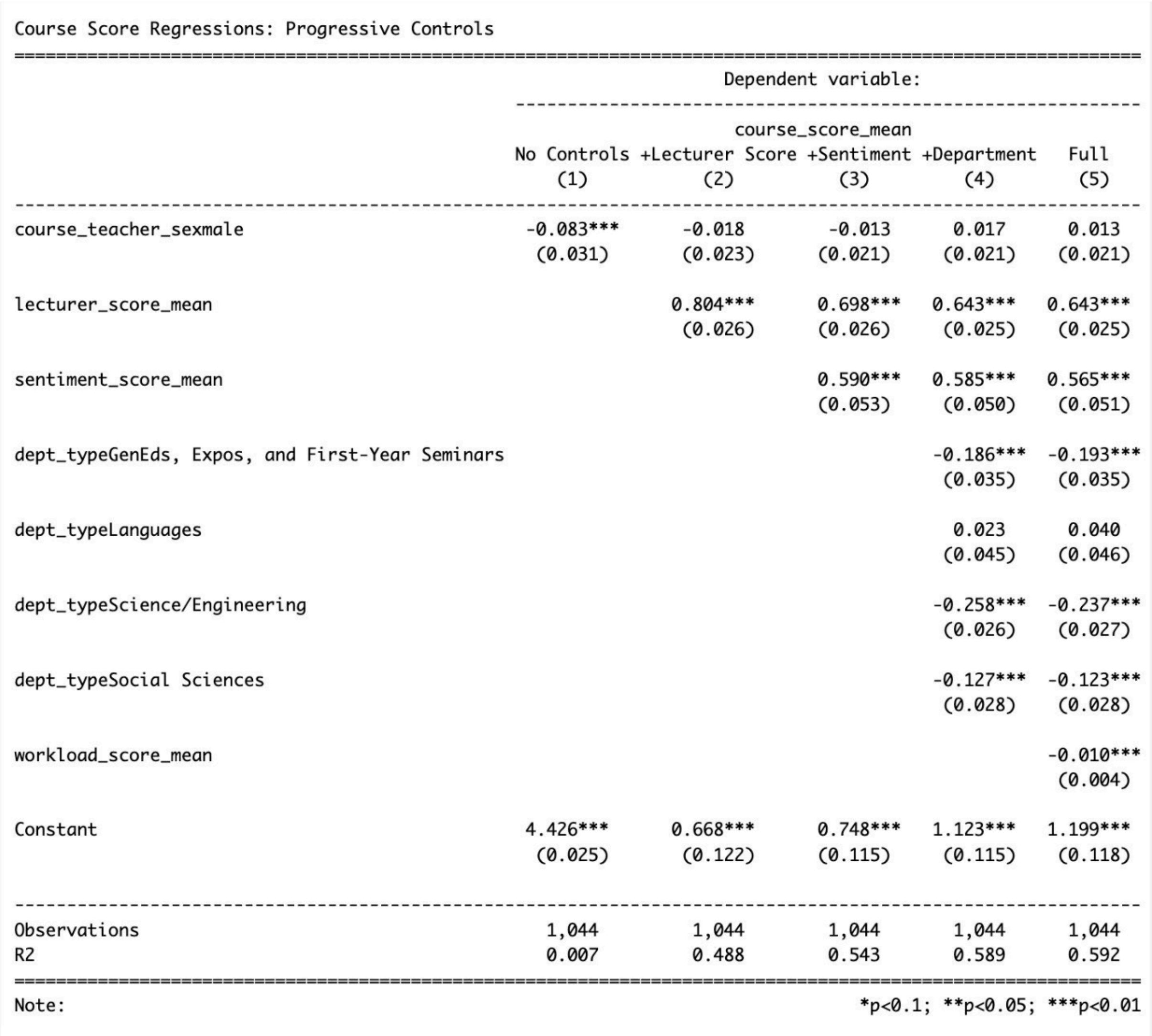
I additionally ran sub-regressions for each primary regression to understand the cumulative effect of each additional control. For the progressive controls analysis of my regression for lecturer score on gender, I find that in a regression with no controls, there seems to be a statistically significant negative relationship between gender and lecturer score, with male instructors receiving slightly lower scores (Fig. 6). However, as more controls are added—department, sentiment score, then course score—and the R squared increases, the negative relationship decreases in magnitude and becomes less statistically significant.



```
Lecturer Score Regressions: Progressive Controls
=============================================================================================================
                                                                     Dependent variable:
                                              ---------------------------------------------------------------
                                                                     lecturer_score_mean
                                              No Controls +Department +Sentiment Score +Course Score   Full
                                                  (1)         (2)           (3)             (4)         (5)
-------------------------------------------------------------------------------------------------------------
course_teacher_sexmale                         -0.082***   -0.053**       -0.041         -0.035*     -0.033
                                                (0.027)     (0.027)       (0.025)         (0.020)    (0.020)

dept_typeGenEds, Expos, and First-Year Seminars            -0.104**      -0.124***         0.035       0.040
                                                            (0.045)       (0.042)         (0.034)    (0.034)

dept_typeLanguages                                          0.040         0.010          -0.008      -0.017
                                                            (0.059)       (0.056)         (0.044)    (0.044)

dept_typeScience/Engineering                               -0.224***     -0.195***         0.035       0.023
                                                            (0.033)       (0.031)         (0.026)    (0.027)

dept_typeSocial Sciences                                   -0.096***     -0.097***         0.016       0.014
                                                            (0.037)       (0.035)         (0.028)    (0.028)

sentiment_score_mean                                                      0.695***         0.077      0.086*
                                                                          (0.058)         (0.051)    (0.052)

course_score_mean                                                                        0.598***   0.601***
                                                                                          (0.023)    (0.024)

workload_score_mean                                                                                  0.006*
                                                                                                     (0.003)

Constant                                       4.676***    4.767***      4.276***        1.959***   1.904***
                                                (0.021)     (0.029)       (0.049)         (0.099)    (0.104)

-------------------------------------------------------------------------------------------------------------
Observations                                     1,044       1,044         1,044           1,044      1,044
R2                                               0.009       0.058         0.173           0.491      0.493
=============================================================================================================
Note:                                                                         *p<0.1; **p<0.05; ***p<0.01
```

Figure 6: Regression Table for Regression of Lecturer Score on Lecturer Gender, with Progressive Controls. Data Source: Harvard QGuide Spring 2025

The full model had no statistically significant difference in the relationship between lecturer score and lecturer gender, suggesting that observed differences in lecturer evaluations between genders are largely explained by course characteristics rather than lecturer gender. This progressive controls analysis further supports the rejection of my hypothesis, as it does not seem that there is gender bias against female lecturers, nor does gender bias seem to explain discrepancies in lecturer scores as much as confounding course characteristics do.

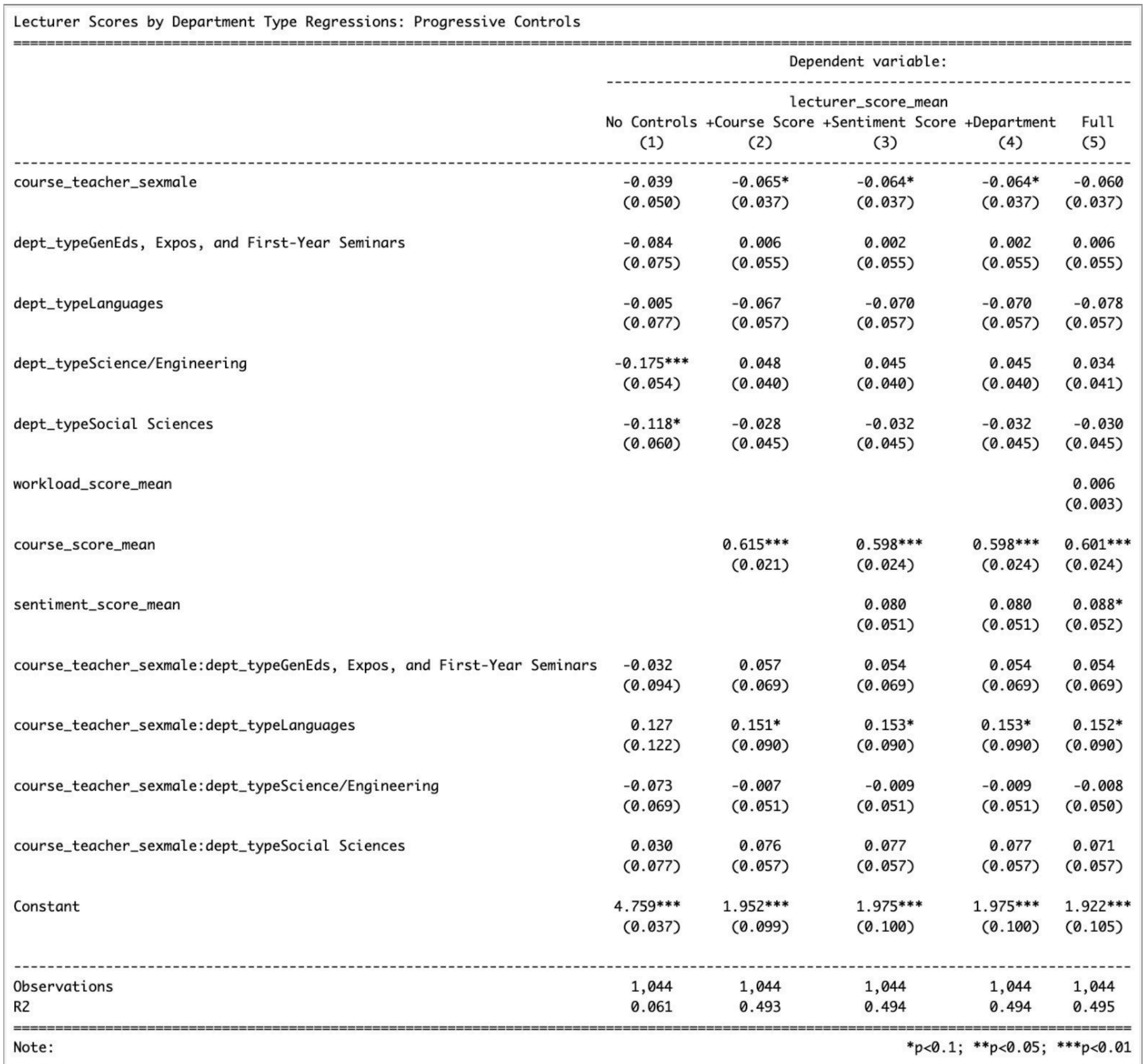


```
Course Score Regressions: Progressive Controls
===============================================================================================================
                                                                   Dependent variable:
                                                 -------------------------------------------------------------
                                                                    course_score_mean
                                                 No Controls +Lecturer Score +Sentiment +Department    Full
                                                     (1)            (2)          (3)         (4)        (5)
---------------------------------------------------------------------------------------------------------------
course_teacher_sexmale                            -0.083***       -0.018       -0.013       0.017      0.013
                                                   (0.031)        (0.023)      (0.021)     (0.021)    (0.021)

lecturer_score_mean                                              0.804***     0.698***    0.643***   0.643***
                                                                  (0.026)      (0.026)     (0.025)    (0.025)

sentiment_score_mean                                                          0.590***    0.585***   0.565***
                                                                               (0.053)     (0.050)    (0.051)

dept_typeGenEds, Expos, and First-Year Seminars                                          -0.186***  -0.193***
                                                                                           (0.035)    (0.035)

dept_typeLanguages                                                                          0.023      0.040
                                                                                           (0.045)    (0.046)

dept_typeScience/Engineering                                                             -0.258***  -0.237***
                                                                                           (0.026)    (0.027)

dept_typeSocial Sciences                                                                 -0.127***  -0.123***
                                                                                           (0.028)    (0.028)

workload_score_mean                                                                                 -0.010***
                                                                                                      (0.004)

Constant                                           4.426***      0.668***     0.748***    1.123***   1.199***
                                                   (0.025)        (0.122)      (0.115)     (0.115)    (0.118)

---------------------------------------------------------------------------------------------------------------
Observations                                        1,044          1,044        1,044       1,044      1,044
R2                                                  0.007          0.488        0.543       0.589      0.592
===============================================================================================================
Note:                                                                          *p<0.1; **p<0.05; ***p<0.01
```

Figure 7: Regression Table for Regression of Course Score on Lecturer Gender, with Progressive Controls. Data Source: Harvard QGuide Spring 2025

I run the same progressive controls analysis on the second regression of course score on lecturer gender (Fig. 7). With no controls, it seems that male lecturers have, on average, a lower course score than female lecturers. However, as the R squared increases, and I control for lecturer score, sentiment score, department, and workload, the magnitude of the relationship decreases and also turns positive, in addition to becoming statistically insignificant. Notably, it seems that lecturer score and course score have a strong correlation, and sentiment score and course score also have a weaker, but still strong correlation, suggesting that these two variables partially explain what the negative and statistically significant coefficient in the regression without controls was picking up on. Similar to the progressive controls analysis for the first regression, course scores do not appear to be statistically associated with gender; instead the course score seems to be driven primarily by perceived instructor quality, student sentiment, department type, and workload. The progressive controls analysis of the lecturer score and course score regressions supports the rejection of my hypothesis, explaining discrepancies in lecturer and course scores as products of course characteristics rather than lecturer gender.

Finally, I also run a progressive controls analysis on the third gender-department interaction-term regression (Fig. 8).



```
Lecturer Scores by Department Type Regressions: Progressive Controls
==============================================================================================================================
                                                                                          Dependent variable:
                                                                      --------------------------------------------------------
                                                                                          lecturer_score_mean
                                                                      No Controls +Course Score +Sentiment Score +Department  Full
                                                                          (1)          (2)            (3)            (4)       (5)
------------------------------------------------------------------------------------------------------------------------------
course_teacher_sexmale                                                  -0.039       -0.065*        -0.064*        -0.064*   -0.060
                                                                        (0.050)      (0.037)        (0.037)        (0.037)   (0.037)

dept_typeGenEds, Expos, and First-Year Seminars                         -0.084        0.006          0.002          0.002     0.006
                                                                        (0.075)      (0.055)        (0.055)        (0.055)   (0.055)

dept_typeLanguages                                                      -0.005       -0.067         -0.070         -0.070    -0.078
                                                                        (0.077)      (0.057)        (0.057)        (0.057)   (0.057)

dept_typeScience/Engineering                                           -0.175***      0.048          0.045          0.045     0.034
                                                                        (0.054)      (0.040)        (0.040)        (0.040)   (0.041)

dept_typeSocial Sciences                                                -0.118*      -0.028         -0.032         -0.032    -0.030
                                                                        (0.060)      (0.045)        (0.045)        (0.045)   (0.045)

workload_score_mean                                                                                                            0.006
                                                                                                                              (0.003)

course_score_mean                                                                     0.615***       0.598***       0.598***  0.601***
                                                                                     (0.021)        (0.024)        (0.024)   (0.024)

sentiment_score_mean                                                                                 0.080          0.080     0.088*
                                                                                                    (0.051)        (0.051)   (0.052)

course_teacher_sexmale:dept_typeGenEds, Expos, and First-Year Seminars  -0.032        0.057          0.054          0.054     0.054
                                                                        (0.094)      (0.069)        (0.069)        (0.069)   (0.069)

course_teacher_sexmale:dept_typeLanguages                                0.127        0.151*         0.153*         0.153*    0.152*
                                                                        (0.122)      (0.090)        (0.090)        (0.090)   (0.090)

course_teacher_sexmale:dept_typeScience/Engineering                     -0.073       -0.007         -0.009         -0.009    -0.008
                                                                        (0.069)      (0.051)        (0.051)        (0.051)   (0.050)

course_teacher_sexmale:dept_typeSocial Sciences                          0.030        0.076          0.077          0.077     0.071
                                                                        (0.077)      (0.057)        (0.057)        (0.057)   (0.057)

Constant                                                                4.759***     1.952***       1.975***       1.975***  1.922***
                                                                        (0.037)      (0.099)        (0.100)        (0.100)   (0.105)

------------------------------------------------------------------------------------------------------------------------------
Observations                                                             1,044        1,044          1,044          1,044     1,044
R2                                                                       0.061        0.493          0.494          0.494     0.495
==============================================================================================================================
Note:                                                                                          *p<0.1; **p<0.05; ***p<0.01
```

Figure 8: Regression Table for Regression of Lecturer Score on Lecturer Gender, Gender-Department Interaction, with Progressive Controls. Data Source: Harvard QGuide Spring 2025

From this progressive controls analysis, it is clear that the difference in the relationship between gender and lecturer score across departments remains insignificant even as more controls are added and the R squared increases. While there is weak evidence of a slightly larger advantage for male lecturers in the Languages department across controls, the magnitude of the relationship is still small and insignificant in the 1-5 scale context. The progressive controls analysis confirms that there is no meaningful difference in the relationship between gender and lecturer score across departments, a rejection of my initial hypothesis.

Conclusion

This study has shown that based on the 2025 spring Harvard QGuide student evaluation, there does not seem to be meaningful gender bias in how students evaluate their lecturers, nor is there a difference in any potential manifestation of gender bias in evaluation across departments. My findings stand in contrast to the existing literature, which has overwhelmingly found evidence of gender bias in student evaluations of teachers. Contrary to Mathisen's findings about Norwegian teenagers, my study suggests that Harvard students may not be similarly moving towards more polarized, right-wing ideologies driven by modern sexism. Rather, due to the recent political climate at Harvard—with several active lawsuits against the right-wing Trump Administration throughout spring 2025—the lack of sexism found in student evaluations might suggest that students at Harvard have been pushed more towards left-wing ideologies in response to these political events; the increasing hostility of the right-wing towards Harvard may have caused Harvard students to become more receptive to progressive ideas of combatting bias and gender equality, perhaps explaining why, contrary to other studies of universities in the literature, Harvard notably displays no significant evidence of gender bias in student evaluations.

There are several limitations with my study and areas for further research. Gender in this dataset was imperfectly coded, relying on scraping lecturers' first names and extrapolating gender using an LLM. While it is fair to say that the sample is still representative of Harvard faculty, the results must be qualified by recognizing the shortcomings of the gender variable. Furthermore, this study does not consider the qualitative data available in the QGuide. The literature on gender bias in student evaluation does suggest that gender bias manifests itself heavily in written comments about instructors, a potential area for future analysis. Finally, because this study only examines attitudes from the spring 2025 semester, the findings are a snapshot in time as opposed to a larger trend. Future research can evaluate the gender bias of the QGuide over time, offering a more robust assessment of how Harvard students' gender bias may change due to political events.

Works Cited

"Data Snapshot: Full-Time Women Faculty and Faculty of Color | AAUP." 2020. https://www.aaup.org/news/data-snapshot-full-time-women-faculty-and-faculty-color (December 10, 2025).

Fincham, Jack E. 2008. "Response Rates and Responsiveness for Surveys, Standards, and the Journal." *American Journal of Pharmaceutical Education* 72(2): 43. doi:10.5688/aj720243.

"'For the Reinvention of Man': How a Conservative Debating Society at Harvard Pushed Women From Its Ranks | News | The Harvard Crimson." 2025. https://www.thecrimson.com/article/2025/12/8/john-adams-society-women/ (December 11, 2025).

Harper, Shaun. "Harvard's First Black Woman President Survived Only Six Months." *Forbes*. https://www.forbes.com/sites/shaunharper/2024/01/04/what-it-means-that-a-black-woman-survived-only-six-months-as-harvards-president/ (December 11, 2025).

Mathisen, Ruben Berge. 2025. "Growing Apart: Ideological Polarization between Teenage Boys and Girls." doi:10.31219/osf.io/7z2va.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. 2017. "Gender Bias in Teaching Evaluations."

Mitchell, Kristina M. W., and Jonathan Martin. 2018. "Gender Bias in Student Evaluations." *PS: Political Science & Politics* 51(03): 648–52. doi:10.1017/S104909651800001X.

Nietzel, Michael T. "Women Continue To Outpace Men In College Enrollment And Graduation." *Forbes*. https://www.forbes.com/sites/michaeltnietzel/2024/08/07/women-continue-to-outpace-men-in-college-enrollment-and-graduation/ (December 10, 2025).

Owen, Ann L., Erica De Bruin, and Stephen Wu. 2025. "Can You Mitigate Gender Bias in Student Evaluations of Teaching? Evaluating Alternative Methods of Soliciting Feedback." *Assessment & Evaluation in Higher Education* 50(3): 442–57. doi:10.1080/02602938.2024.2407927.

Saygin, Perihan O., and Xi Zhang. 2025. “Gender Gap in Teaching Evaluations and Its Effect on Course Enrollments.” *Economics of Education Review* 104: 102617. doi:10.1016/j.econedurev.2024.102617.

Appendix 1: Documentation of Interaction with Codex by OpenAI, an LLM Code Assistant

Codex is a coding-helper tool that integrates directly into a user's local computer terminal. Codex was used to generate a Python script that scraped lecturers' first names from the QGuide links in the dataset. Codex also produced code to cross reference lecturers' first names with a database of gender and names to add the lecturers' gender to each observation of the dataset.

Query 1: Write me a python script that goes through each row in 2025springQ.csv and adds the professor's first name to the file. The last name is in the "course_teacher" column, the department and course number is in course_code. The name can be either searched from Harvard staff directory, or parsed from the URL in the link field. These are all Harvard courses and Harvard staff.

Query 2: Plug in a gender-guess library and add a sex column to the csv, skip the first name extraction if it already exists.