W241 Final Project Paper

Elizabeth Khan, Estrella Ndrianasy, Chandni Shah, Michelle Shen, Catherine Tsai

Contents

Background

Female instructors face challenges with gender stereotypes that can impact their future earning potentials and career growth. Specifically for women in academia, numerous studies have demonstrated that students treat and evaluate female professors differently than male professors. Female professors are also often held to higher standards and subjected to a greater number of demands and requests from their students.[2] In this experiment, we seek to understand if gender bias alters students' perception of an instructor's quality of programming and technical instruction.

Research Question

Does an instructor's perceived gender influence the perceived quality of instruction?

Hypothesis

We hypothesize that the treatment of changing perceived gender of the instructor will impact the measured instructor ratings from students due to historical evidence of gender bias in academia favoring men's performance, grant and award-winning potential, resource allocation, and tenure that continues to reinforce gender wage gaps and lead to better career outcomes for men. Specifically, in an academic learning setting, we suspect that implicit gender biases against women can be observed when evaluating instructor performance by exposing the treatment group to instructors they perceive as female.

Experiment Design

Our experiment featured a randomized, between-subjects design with a 2x4 blocked design on participants' gender and age groups to ensure equal treatment distribution among each of the blocked populations. The sample design consisted of 112 participants equally split between males (n = 56) and females (n = 56). Prior to conducting the experiment, a statistical power test was completed to understand if having a sample size of at least 112 participants (56 subjects in the treatment and control groups) would be sufficient to observe the treatment effect (if indeed there is one). The treatment effect in this simulation was estimated from the average outcome from the Gender Bias in Student Evaluations Study by Kristina Mitchell and Johnathan Martin. The statistical power test demonstrated that 95% of all potential random assignments will reject the null hypothesis if indeed there is a treatment effect. Therefore, this suggests that a sample size of at least 112 is enough to detect the treatment effect.

Methods

A video featuring a Python programming informational slideshow was created. Three (3) men and three women volunteers recorded voice overs narrating the video using the same script. Volunteers were given timestamps to standardize pacing of the 6 resulting videos. Neither videos nor images of volunteers were incorporated into the final videos. A questionnaire containing thirteen (13) survey questions (see Appendix A) was generated to collect information on two main outcomes. First, the subject is asked five (5) primary outcome questions to collect information on how a subject perceived their instructor from the video. Primary outcome questions use a 5-point Likert Scale for subjects to rate the quality of the instructor's performance in categories such as overall instructor effectiveness, competence/knowledge in subject matter, instructor enthusiasm and professionalism. One (1) attention check question is administered at this time. Next, the subject is asked four (4) secondary objective outcome questions to test video content retention among respondents. Questions are presented in multiple-choice format with one correct response. Lastly, relevant demographic information from each subject is collected with three (3) final questions.

Table 1: Respondents by Age Group and Gender

Characteristic	**Overall**, N = 221	**Female**, N = 142	**Male**, N = 79
assignment			
Control	108 (49%)	68 (48%)	40 (51%)
Treatment	113 (51%)	74 (52%)	39 (49%)
Age			
20-30	44 (20%)	25 (18%)	19 (24%)
30-40	82 (37%)	56 (39%)	26 (33%)
40-50	48 (22%)	32 (23%)	16 (20%)
50+	47 (21%)	29 (20%)	18 (23%)

Six Qualtrics surveys were generated. Each Qualtrics survey contained only one of the six videos with either a male or a female voice over, followed by the questionnaire. A 10 second timer was added to each page of the survey.

Recruitment Process

A social science subject-recruiting company called Survey Swap was used to identify subjects, administer the Qualtrics video and survey, and record subject responses. To qualify for the survey, we stipulated subjects must be located in the United States, be over the age of 20, and identify as a native English speaker. The goal of this criteria requirement was to allow us to capture treatment effects for adults in the United States. Subjects who qualified for the survey were randomly redirected to one of the six Qualtrics surveys, where they were asked to view the video. After viewing the video, subjects were asked to fill out the survey questions. Survey responses from only subjects that correctly answered the attention check question responses were recorded.

We defined the control group as consisting of subjects who watched an instructional video voiced using a male voiceover and the treatment group as consisting of subjects who were treated with the same video voiced using a female instead of a male voiceover. To control for cadence, pitch, tone, and other characteristics that vary among both men's and women's voices, we used a total of six voiceovers from three men and three women.

Data Preperation

Subject Demographics

Data from 222 total subjects was received from Survey Swap. Of those subjects, 49% were assigned to the control group and 51% were assigned to the treatment group. In the tables below you can observe the full summary of subjects within each gender and age block, along with education level.

Randomization

While blocked randomization was completed through the distribution of surveys by Survey Swap, we conducted our own randomization check upon receiving the data. To conduct this check, we utilized an F-Test in R to test if subjects were equally assigned to the control and treatment group based on their age and gender blocks. The p-value for this test was not statistically significant at 0.795 and therefore we fail to reject the null hypothesis. The results of our test suggest covariate balance within each block and that the blocks were successfully randomized.

Table 2: Respondents by Education and Treatment Assignment

Characteristic	**Overall**, N = 221	**Control**, N = 108	**Treatment**, N = 113
Gender			
Female	142 (64%)	68 (63%)	74 (65%)
Male	79 (36%)	40 (37%)	39 (35%)
Education			
Associates degree	27 (12%)	16 (15%)	11 (9.7%)
Bachelors degree	43 (19%)	21 (19%)	22 (19%)
High school diploma	66 (30%)	33 (31%)	33 (29%)
Less than High school	7 (3.2%)	1 (0.9%)	6 (5.3%)
Masters degree	11 (5.0%)	5 (4.6%)	6 (5.3%)
Some College No degree	67 (30%)	32 (30%)	35 (31%)

Pre-Treatment Covariates

We hypothesized gender, age, and level of education of subjects may influence the outcome measure. To account for these covariates, a covariate balance check was conducted to evaluate if the distributions of the covariates were similar between the treatment and control groups. Findings suggested there is no imbalance between pre-treatment covariates, as none of the p-values were statistically significant.

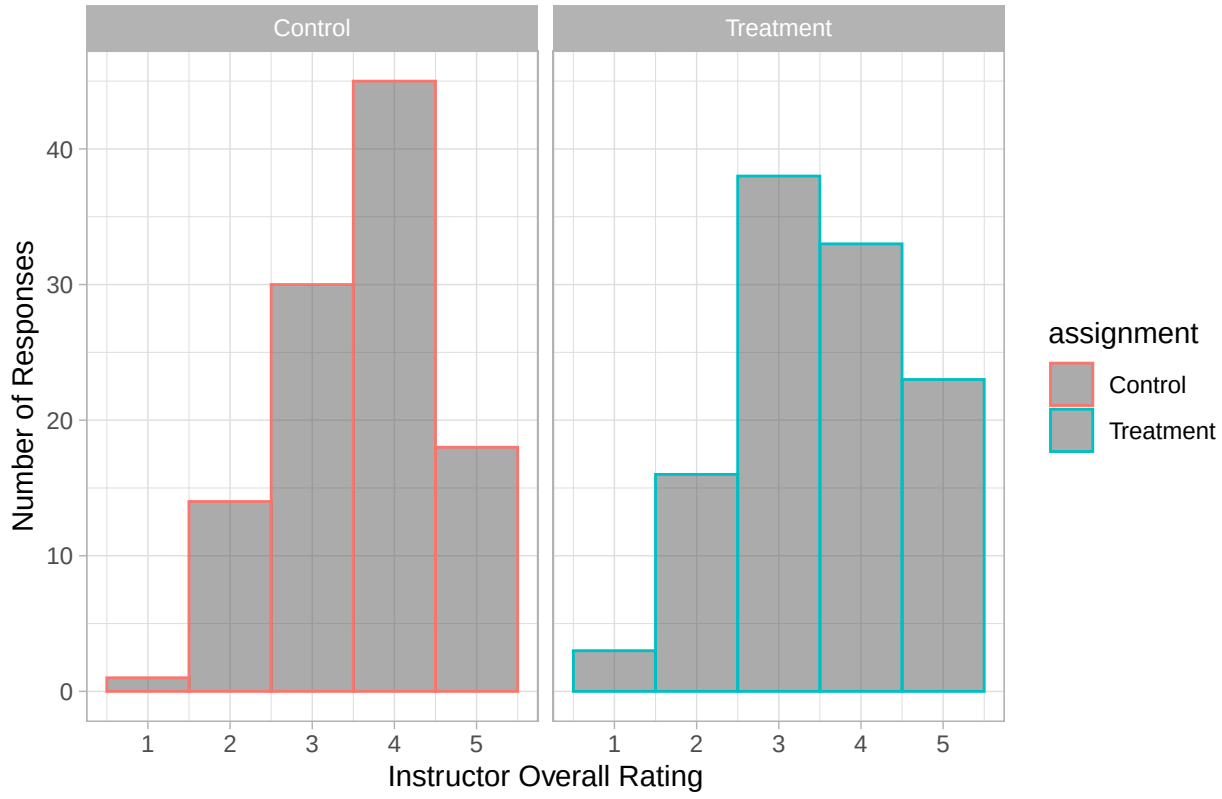
Table 3: Table 3: Covariate Balance Test

	Control (N = 108)	Treatment (N = 113)	% - Control	% - Treatment	t-test (Diff in Means)
Gender					
Male	40 (37.04%)	39 (34.51%)	0.37	0.35	0.697
Female	68 (62.96%)	74 (65.49%)	0.63	0.65	0.697
Age					
20-30	23 (21.30%)	21 (18.58%)	0.21	0.19	0.616
30-40	39 (36.11%)	43 (38.05%)	0.36	0.38	0.766
40-50	26 (24.07%)	22 (19.47%)	0.24	0.19	0.41
50+	20 (18.52%)	27 (23.89%)	0.19	0.24	0.33
Education					
Less than High school	1 (0.93%)	6 (5.31%)	0.01	0.05	NA
High school diploma	33 (30.56%)	33 (29.20%)	0.31	0.29	0.827
Some College No degree	32 (29.63%)	35 (30.97%)	0.3	0.31	0.829
Associates degree	16 (14.81%)	11 (9.73%)	0.15	0.1	0.253
Bachelors degree	21 (19.44%)	22 (19.47%)	0.19	0.19	0.996
Masters degree	5 (4.63%)	6 (5.31%)	0.05	0.05	0.817

Results

While data from only 112 subjects was requested, Survey Swap provided responses from 222 total subjects who passed the attention check. Our analysis explores both the overall instructor rating (an average of the five ratings) and each instructor rating separately. From the histogram below, we can observe slight differences in the distribution of ratings within the treatment and control groups for the overall instructor rating. While the distribution of both groups are slightly skewed to the left, we observe the majority of subjects in the control group rated the instructor 4 out of 5 and the majority of subjects in the treatment group rated the instructor 3 out of 5. While the distributions of the assignment groups differ, we can see in the box plot below the mean ratings between groups are similar.

Figure 1: Instructor Overall Rating by Treatment and Control




```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(col_plot)` instead of `col_plot` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

Figure 2: Overall Instructor Ratings by Treatment Assignment

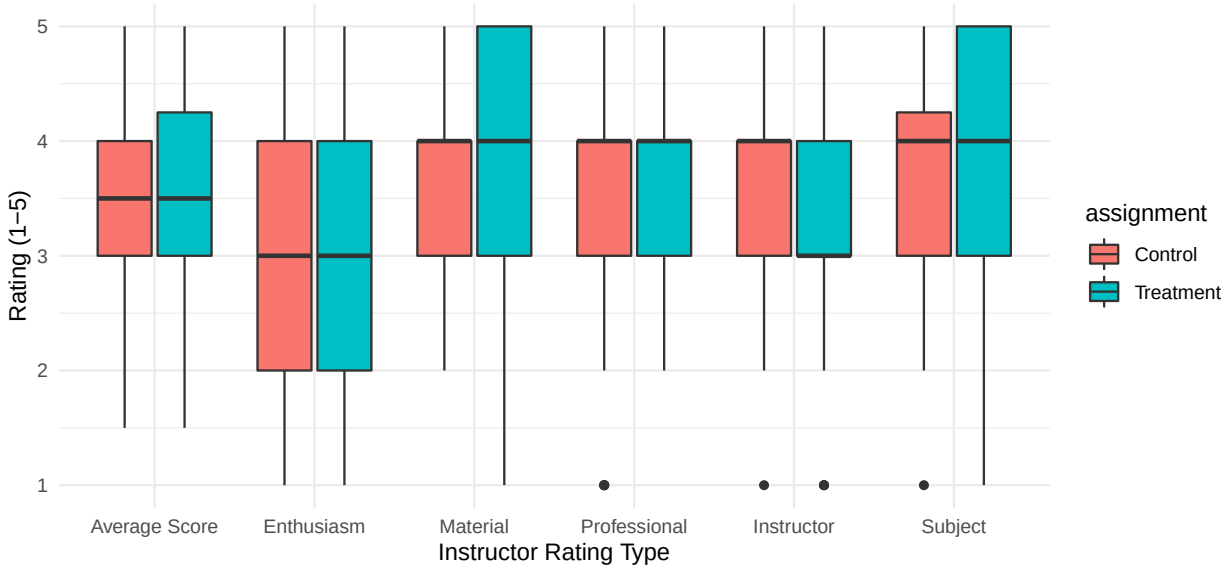


Table 4: Baseline Model Results

	Dependent variable:				
	Average Rating	Instructor	Subject	Material	Enthusiasm
	(1)	(2)	(3)	(4)	(5)
Male Instructor Video	0.122	-0.097	0.068	0.176	0.364**
	(0.114)	(0.135)	(0.122)	(0.137)	(0.152)
Constant	3.450***	3.600***	3.860***	3.630***	2.780***
	(0.082)	(0.096)	(0.087)	(0.098)	(0.109)
Observations	221	221	221	221	221
R^2	0.005	0.002	0.001	0.007	0.025
Adjusted R^2	0.001	-0.002	-0.003	0.003	0.021
Residual Std. Error (df = 219)	0.849	1.000	0.907	1.020	1.130
F Statistic (df = 1; 219)	1.140	0.521	0.311	1.640	5.710**
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 5: Outcome with Blocks

	Dependent variable:				
	Average Rating	Instructor	Subject	Material	Enthusiasm
	(1)	(2)	(3)	(4)	(5)
Age: 30-40	-0.118	-0.109	-0.212	-0.098	0.043
	(0.159)	(0.188)	(0.170)	(0.192)	(0.208)
Age: 40-50	-0.094	0.097	-0.049	0.002	-0.306
	(0.177)	(0.210)	(0.190)	(0.213)	(0.232)
Age: 50+	-0.131	0.003	0.064	-0.035	-0.144
	(0.178)	(0.211)	(0.191)	(0.215)	(0.233)
Male	0.229*	0.237*	0.108	0.220	0.479***
	(0.120)	(0.141)	(0.128)	(0.144)	(0.156)
Male Instructor Video	0.133	-0.085	0.069	0.185	0.369**
	(0.115)	(0.135)	(0.123)	(0.138)	(0.150)
Constant	3.450***	3.530***	3.900***	3.590***	2.680***
	(0.149)	(0.176)	(0.159)	(0.179)	(0.195)
Observations	221	221	221	221	221
R^2	0.026	0.022	0.021	0.021	0.081
Adjusted R^2	0.003	-0.001	-0.002	-0.002	0.059
Residual Std. Error (df = 215)	0.848	1.000	0.907	1.020	1.110
F Statistic (df = 5; 215)	1.150	0.962	0.913	0.902	3.770***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 6: Interaction between Gender and Treatment

	Dependent variable:				
	Average Rating	Instructor	Subject	Material	Enthusiasm
	(1)	(2)	(3)	(4)	(5)
Age: 30-40	-0.116	-0.108	-0.210	-0.092	0.041
	(0.160)	(0.189)	(0.171)	(0.191)	(0.209)
Age: 40-50	-0.087	0.100	-0.040	0.029	-0.313
	(0.179)	(0.211)	(0.191)	(0.214)	(0.233)
Age: 50+	-0.129	0.004	0.067	-0.026	-0.147
	(0.179)	(0.211)	(0.191)	(0.214)	(0.234)
Male	0.180	0.219	0.042	0.023	0.536**
	(0.170)	(0.201)	(0.181)	(0.203)	(0.222)
Male Instructor Video	0.098	-0.097	0.022	0.045	0.409**
	(0.143)	(0.169)	(0.153)	(0.172)	(0.187)
Male:Male Instructor Video	0.097	0.035	0.132	0.393	-0.113
	(0.240)	(0.283)	(0.256)	(0.287)	(0.314)
Constant	3.470***	3.540***	3.920***	3.650***	2.670***
	(0.154)	(0.182)	(0.165)	(0.185)	(0.201)
Observations	221	221	221	221	221
R^2	0.027	0.022	0.022	0.029	0.081
Adjusted R^2	-0.001	-0.005	-0.005	0.002	0.055
Residual Std. Error (df = 214)	0.850	1.000	0.909	1.020	1.110
F Statistic (df = 6; 214)	0.981	0.800	0.802	1.070	3.150***
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 7: Interaction between Gender and Treatment

	Dependent variable:		
	Quiz Score		
	(1)	(2)	(3)
Age: 30-40		-0.032	-0.032
		(0.046)	(0.046)
Age: 40-50		0.032	0.031
		(0.051)	(0.052)
Age: 50+		0.095*	0.095*
		(0.052)	(0.052)
Male		-0.052	-0.045
		(0.035)	(0.049)
Male Instructor Video	-0.020	-0.024	-0.019
	(0.034)	(0.033)	(0.042)
Male:Male Instructor Video			-0.013
			(0.070)
Constant	0.449***	0.454***	0.452***
	(0.024)	(0.043)	(0.045)
Observations	221	221	221
R^2	0.002	0.047	0.048
Adjusted R^2	-0.003	0.025	0.021
Residual Std. Error	0.250 (df = 219)	0.246 (df = 215)	0.247 (df = 214)
F Statistic	0.350 (df = 1; 219)	2.140* (df = 5; 215)	1.780 (df = 6; 214)
Note:			*p<0.1; **p<0.05; ***p<0.01

Analysis

Discussion

Conclusion

Limitations and Future Enhancements