Examining the Effect of Stereotype Threat on Academic Performance

Tiffany Cheng, Matt Kinkley, Prachi Varma

August 8, 2022

Abstract

Does asking for a subject's gender and increasing the pressure affect their performance on an academic exam? There is existing experimental research to suggest that making subjects aware of their own race prior to an exam leads to worse results for black students when compared to white students. There are also additional studies that either make it clear to subjects that the exam shows gender differences or that the exam does not show gender differences, and female students perform worse than male students in the former scenario. Both of these examples show that groups with negative stereotypes surrounding academic achievement tend to perform worse when given subtle reminders of their identity. The intervention we chose for this experiment is as follows: prior to having the subjects answer SAT math questions, we ask for their gender (female, male, or non-binary) and tell them that the exam measures intellectual ability through a series of math questions. This intervention aims to bring the subject's gender to the forefront and specifically focuses on math questions since females are generally stereotyped to be worse at math than males. One group will receive this intervention and the other group will proceed to the math questions without seeing the gender question or the pressure-inducing sentence. Our findings suggest that there is no heterogeneous treatment effect from our intervention, meaning women who were shown the pressure-inducing statement and asked their gender before taking the math exam did not perform any worse on average than men who received the same treatment. We actually found that the subjects in the treatment group, both men and women, performed better than those who did not receive the experimental intervention. We also found that those in the treatment group were less likely to attrit from the experiment. These results are peculiar as it is possible that our treatment effect was actually motivating rather than stress inducing.

Research Question

Does asking for a subject's gender and increasing the pressure affect their performance on an academic exam?

Through this question we are looking to analyze if stereotype threat, specifically related to gender, will be induced in our subjects when answering SAT level math exam questions. Previous research suggests that STEM questions have a particularly powerful stereotype effect related to gender.

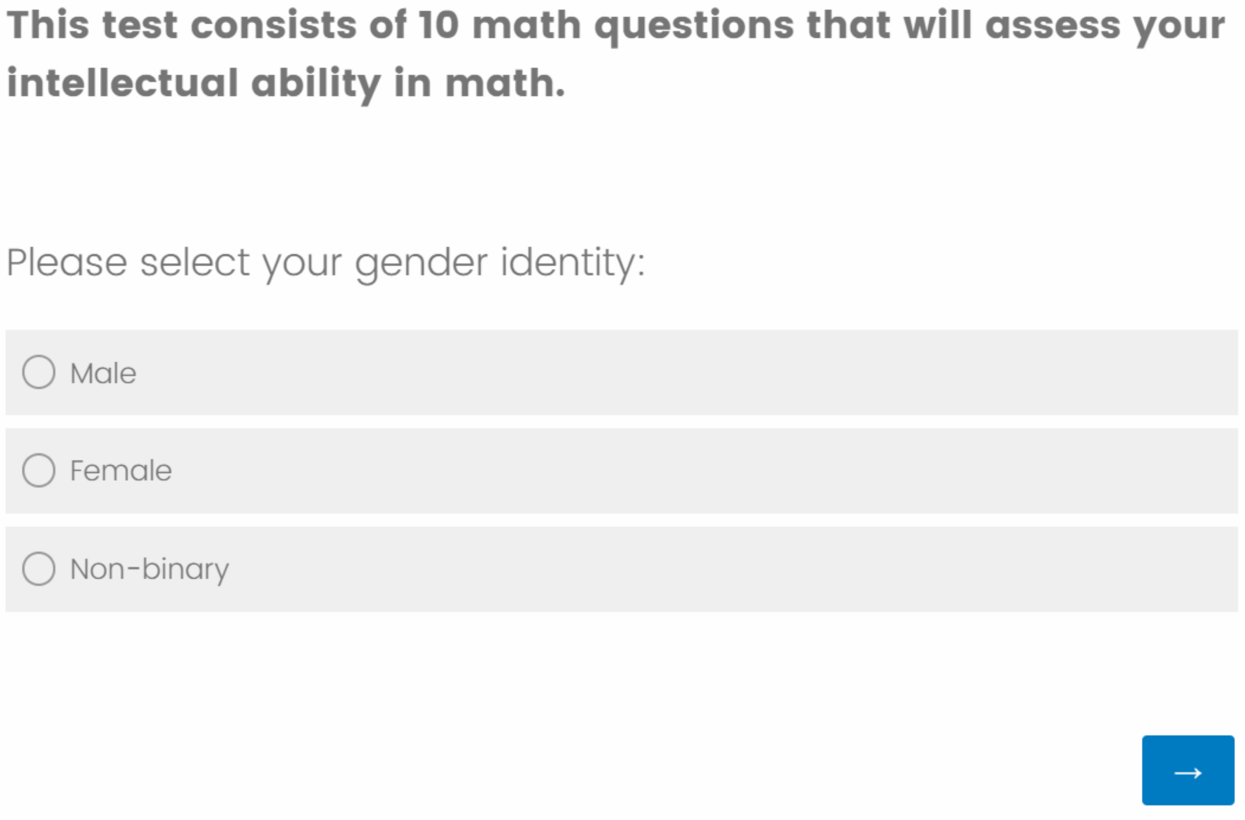
Hypothesis

Stereotype threat is being at risk of confirming a negative stereotype about one's own group. It may manifest when negative stereotypes about one's group performance are reinforced or even hinted at, suggesting apprehension or pressure brought to the front of one's mind when performing a task. For example, (Spencer, 1999) found that women asked to complete math problems performed substantially worse than equally qualified male counterparts when the test was described as producing gender differences. Whereas among women who were told the test resulted in equal scores among genders, there were no differences in scores. Additionally, (Steele, 1995) performed a similar study among White and African-Americans on difficult verbal questions. When the test was described as being an accurate diagnostic of intellectual ability, blacks performed substantially worse than their white counterparts as compared to those who were told the test was merely an exercise and not indicative of intelligence. Both of these experiments utilized a high stereotype-threat induced treatment group that was exposed to a disclaimer about the test's findings or abilities to highlight the prevailing negative stereotype, compared to one where any such threat was reduced. Also, each study's participants were very motivated to complete their tests accurately.

In line with these studies, we believe that being shown negative stereotype-reinforcing language about a group's performance in certain tasks will impact those members' outcomes in those tasks. Therefore, there should be a heterogeneous treatment effect, where women who are exposed to the stress inducing statement and asked to supply their gender before taking the exam should perform worse than those in the treatment group whereas men who were exposed to the same intervention should not display much change in their scores on average when compared to the control group.

Treatment

Our treatment consisted of two components: a stereotype-inducing sentence followed by a question asking for the subject's gender identification. The treatment was presented to the participant before they began the test and a screenshot of the treatment is shown below. The purpose of this was to induce the stereotype threat before the questions began so it was at the forefront of the subject's mind, and we could attribute the resulting scores to be caused by this question. Meanwhile the control group simply saw the following message before they began the test: "This test consists of 10 math questions" and they were asked for their gender at the end of the test, along with the demographic covariate questions which were asked of both groups.



This test consists of 10 math questions that will assess your intellectual ability in math.

Please select your gender identity:

○ Male

○ Female

○ Non-binary

Figure 1: Screenshot of Treatment

Measurement Units

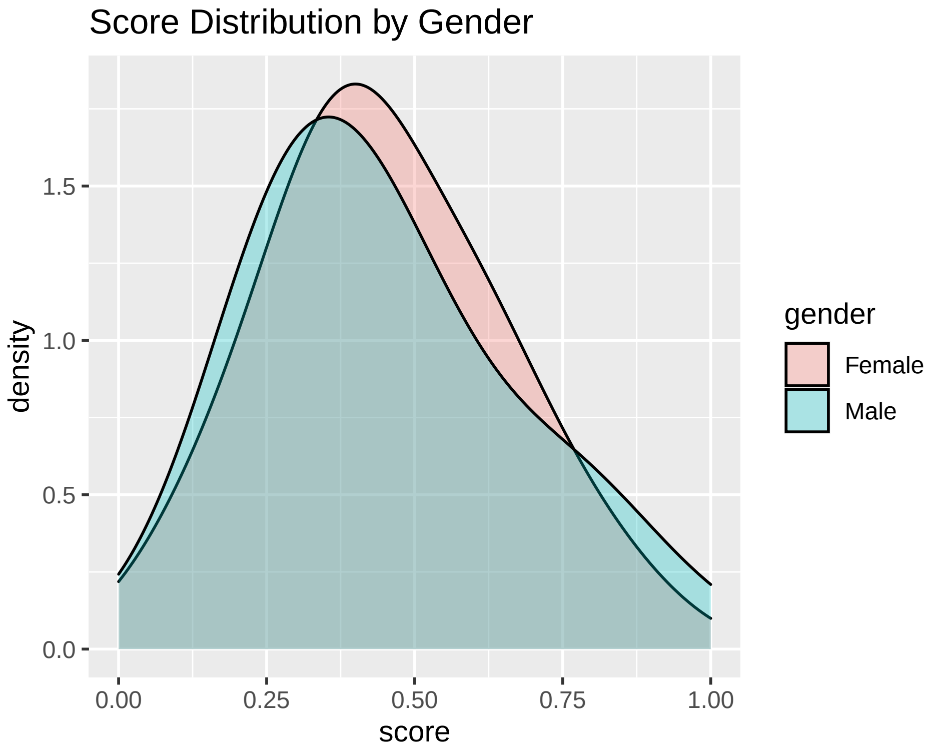
The experiment was conducted on Amazon Web Services' Mechanical Turk workers. MTurk allows anyone with an internet connection to access and complete online assignments or surveys in return for money. Initially we conducted a pilot run of data collection, to estimate how many responses we could expect to get, and how reliable the results might be. We found that there was a relatively even distribution of men and women, which was suitable for our analysis. However there was a significant percentage (~25%) of low-effort responses, which were either all zeroes, or all 10/10s. We discovered that the answers to the math questions were easily found online, which explained the perfect scores achieved with little time taken.

For our final data collection, we adjusted the question answers so they could not be found online. We additionally advertised the survey on MTurk to only "master" workers, who are people that have been deemed by AWS to provide consistently diligent task completions. We gathered 142 completed responses over one week, which were reimbursed to the participants at $0.80 (plus an additional fee to AWS). While the platform did provide the ability to filter on worker demographics, we decided it was not worth the additional cost as our gender distribution was already even from the pilot study. Ultimately our final data collection had fewer low effort responses due to incorporating custom questions and "master" workers.

We had hypothesized that motivation to complete the test accurately was necessary to see a treatment effect. For example, we estimated that the treatment effect resulting in lower scores would be highest in situations where the subject is under a pressurized, competitive atmosphere. In an effort to reproduce these environments and obtain a sample from people who are extremely motivated, we had hoped to pursue additional recruitment avenues through: an SAT preparatory center where local high school students attended, and a Reddit forum dedicated to SAT help. Unfortunately we were unable to move forward with the tutoring center, and the Reddit forum rules did not allow for us to post there.

Outcome Measures

The outcome measurement of our experiment is a score from 0 to 10 out of 10 correct questions answered, expressed as a percentage score between 0 and 1. Our expectation is that with increased pressure and distraction due to stereotype threat, subjects would achieve lower scores. The 10 questions provided on the survey were SAT math questions obtained from a prep book, ranging in 5 levels from very easy to very difficult. The number of questions on our survey that ranged from least to most difficult was 1/1/3/3/2. Concluding our experiment, we found that both men and women had similar, normal, score distributions around 0.4. This instilled confidence in our test being sufficiently difficult, as prior research indicated that a difficult assessment is necessary to see a treatment effect. Seeing a normal distribution with a wide range of scores will also make it easier to observe any treatment effects, as we do not see many scores grouped around 0 or 1.



Experiment Design

The survey was provided on Amazon MTurk displayed as:

> Answer 10 SAT math questions: Help us gather data on math proficiency.
>
> Please read each question and answer to the best of your ability.

When subjects clicked to partake in the task, they received the Qualtrics link to our survey. The link was randomized to provide either the treatment or control survey, the difference between the two being the method in which gender was asked for. The order of the math questions was randomized for each test. After completing the test section, subjects in both groups were then asked for their age, race, education level, income, (and gender in the case of the control group). Finally, subjects received a code to verify their completion and get reimbursed through Amazon. They did not receive a final score or breakdown of correct responses. The results were aggregated and downloaded from Qualtrics once the data collection period was over, and we had gathered 142 responses.

Our experiment follows a randomized post-test observation design. The treatment and control groups are randomized through the Qualtrics link to a 50/50 ratio. The treatment of asking for gender associated with the stereotype-reinforcing language was applied in only the treatment surveys. Then, we conducted a single post-test observation of all responses.

Randomization

To implement our randomization, we used Qualtrics' Randomizer element in the Survey Flow to randomly assign subjects as they entered the survey to either the treatment or the control branch. We also aimed to evenly present treatment and control branches through a setting in the Randomizer element so that we would get roughly even numbers of subjects in treatment and control.

To check for whether the randomization worked, we conducted a covariate balance check by regressing the treatment variable on each of our five covariates (`gender`, `age`, `race`, `education`, `income`). Since we used a logistic regression, we performed a likelihood-ratio test on all covariates and found that no variables were statistically significant at the $\alpha = 0.05$ significance level. Therefore, we can conclude that there are no covariate differences between treatment and control.

Table 1: Likelihood-Ratio Tests for all Explanatory Variables

	LR Chisq	Df	Pr(>Chisq)
gender	1.3702616	1	0.2417669
age	0.0635946	1	0.8009022
race	2.3091473	5	0.8049224
education	10.3190994	7	0.1712000
income	12.0605696	8	0.1485218

In looking at side-by-side charts of our covariates split up by treatment assignment (available in Appendix A), we found that income and education have the most different distributions, although the general trend still remains the same between the two groups. Please note, this covariate data come from subjects that completed the test. From the results of the statistical test and from the charts, we can conclude that we have successfully randomized our subjects into treatment and control groups.

Experiment Flow

There were a total of 245 subjects that participated in our study, 125 of whom were in treatment and 120 of whom were in control. Of those subjects, we were able to measure treatment outcomes for 81 subjects and we were able to measure control outcomes for 61 subjects. We also included a third gender option of "non-binary" in our survey, but we did not see any instances of non-binary subjects in our study.

We saw attrition occur and found that 35% of the treatment group did not complete the test and 49% of the control group did not complete the test. For the attriters in the treatment group, 52% were female and 48% were male (p-value from t-test: 0.8312). Unfortunately, we cannot break down the attrition in the control group by gender because subjects were asked to select their gender identification after they completed the test. In addition, we structured our survey such that all our covariate questions came at the very end (age, race, education, income) for both groups, so we are not able to definitively say if there are patterns in attrition in one group. If we had collected covariate data prior to administering the test, we could run a regression model to predict attrition using the covariate data. If all covariates were not statistically significant, we would conclude that there was random attrition; however, if one or more covariates were statistically significant, we would conclude that there is a certain group of subjects that is more likely to not complete the survey. With this knowledge, we could remove would-be non-compliers from our treatment group based on their covariate characteristics, and calculate the treatment effect among compliers in both groups only.

There is evidence of differential attrition, since the difference in attrition between treatment and control is statistically significant at the $\alpha = 0.05$ significance level (p-value from t-test: 0.0371). One theory we have as to why there was differential attrition is that perhaps the treatment group found the stereotype-inducing sentence motivating and it actually encouraged them to take and to complete the test. If this were the case, we would overestimate our treatment effect. This is because we assume that participants who were motivated by the treatment question, would also be motivated to complete the test accurately. Conversely, those who failed to complete the exam, could be assumed to have lower scores as they got frustrated with the difficulty of the test. We also made sure to randomize the order in which questions were presented to participants, so we cannot identify which question made certain participants leave the test. In addition, Qualtrics does not record a participant's responses until they click the blue "Next" button, so we were not able to see how much progress someone made before they left that survey page. Perhaps a lesson learned for future survey design is to ask as many covariate questions as possible first before administering the actual survey.

EDA

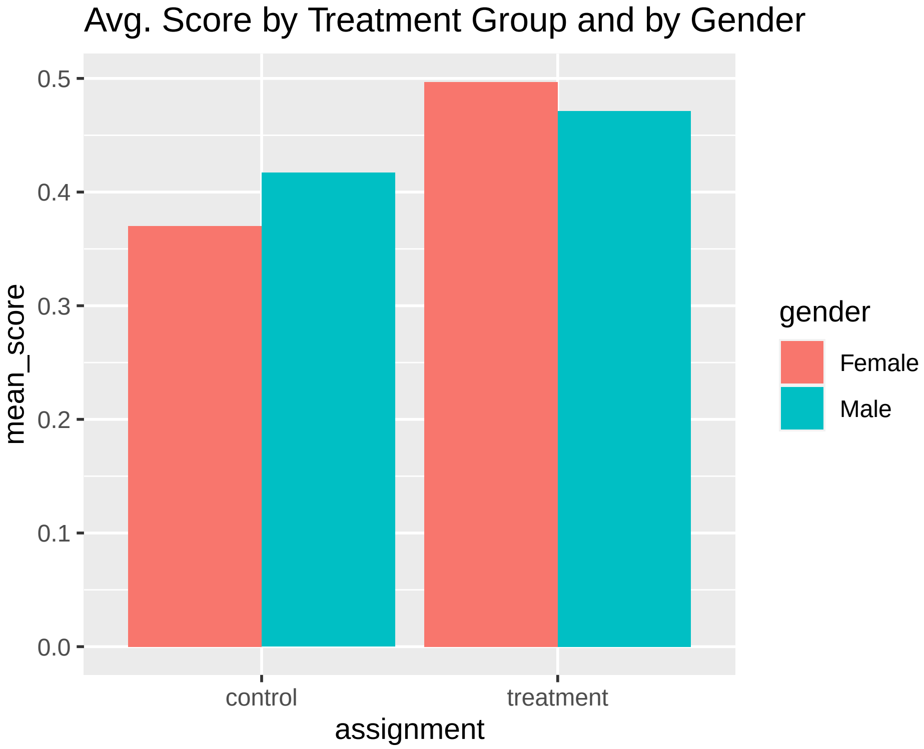
Before we built any models, we performed some exploratory data analysis to better understand our covariates. Additional tables and graphs beyond those shown below are included in Appendix B. The table below shows the breakdown between gender and treatment assignment. Roughly one third of our sample size of 142 is female and the rest is male. This proportion also holds for the treatment and control groups as well.

Table 2: Gender and Treatment Assignment Contingency Table

	Treatment	Control	Total
Female	32	20	52
Male	49	41	90
Total	81	61	142

For the rest of the covariates besides gender, we found that most of our subjects are between 30-50 years old, they also overwhelmingly identify as White, they tend to earn less than $100,000 annually, and most of them have a Bachelor's degree.

The chart below replicates the format we typically saw in prior research conducted on stereotype threat and aims to show how subjects performed on our test. We separated our dataset by gender and treatment assignment and computed the average score within each group to compare within and between treatment assignment and gender. From our data, both males and females performed better in treatment when compared to control, but females saw the biggest performance increase going from control to treatment.



Analysis and Results

Our outcome of interest is score, which ranges from 0 to 1 and measures the proportion of answers the participant answered correctly. The variable treatment is a binary indicator for whether the subject was in the treatment group (1) or the control group (0). There are five additional covariates used in the analysis and they are gender (Female or Male), age (numeric), race (factor with 6 levels), education (factor with 8 levels), and income (factor with 9 levels).

While we were fitting our models, we realized that there was only one subject with a race of "Other" and only one subject with an income of $175,000-$199,999. We had to remove those two rows of data because we could not fit a model with race and income when they were included in the dataset.

Our first model ("Limited"), estimates a single treatment-control contrast. From the model output, our treatment is statistically significant and has a positive effect on subjects' performance. Specifically, we see that the treatment group scored 48%, whereas the control group scored 40%, on average. When we add our covariates (gender, age, race, education, income) to this model ("Limited w/ Controls"), we find that our treatment effect and standard error remain unchanged and the treatment estimate is still statistically significant. The intercept estimate has changed because the variation in the data has now been spread out between more variables.

Our second set of models aims to test for the existence of heterogeneous treatment effects. From previous research, we had hypothesized that there would be a negative treatment effect for females and no treatment effect for males. The third model ("HTE") explores this hypothesis with main effects for treatment and gender and an interaction term between the two variables. Since the interaction term is not statistically significant, we do not see evidence of a heterogeneous treatment effect. From the coefficients, it seems that females performed better than males in treatment and females performed worse than males in control, which is consistent with our chart from the EDA section. Keeping in mind the insignificance, we hesitate to put too much weight behind the interpretation of these estimates. Our final model ("HTE w/ Controls) adds covariates onto our third model (age, race, education, income), but we do not see any increase in precision to our estimates in our third model.

Before we designed our test and collected data, we performed a power analysis to determine our sample size under various heterogeneous treatment effect size assumptions. Although we initially assumed a negative treatment effect for females but observed a positive treatment effect, our power analysis results still hold. For a sample size of 140 and a heterogeneous treatment effect of 0.07, our calculated power is at least 0.7. Appendix C has more information on our power analysis that includes a chart summarizing our findings.

The table below displays a summary of all four models with robust standard errors in parenthesis.

Table 3: Model Results

	Dependent variable:			
	Test Score (as percentage)			
	Limited	Limited w/ Controls	HTE	HTE w/ Controls
	(1)	(2)	(3)	(4)
Treatment (Yes)	0.08** (0.04)	0.08** (0.04)	0.06 (0.05)	0.07 (0.05)
Gender (F)		−0.04 (0.04)	−0.04 (0.06)	−0.06 (0.06)
Age		0.004 (0.003)		0.004 (0.003)
Native American		−0.09 (0.12)		−0.10 (0.12)
Asian		−0.03 (0.06)		−0.03 (0.06)
African American or Black		0.08 (0.06)		0.08 (0.06)
Hispanic or Latinx		−0.11 (0.10)		−0.11 (0.10)
High school graduate		0.15 (0.09)		0.14 (0.09)
Some college credit, no degree		0.20** (0.08)		0.20** (0.08)
Trade/technical/vocational		0.13 (0.11)		0.14 (0.11)
Associate		0.19* (0.10)		0.19* (0.10)
Bachelor		0.22*** (0.08)		0.22*** (0.08)
Master		0.26** (0.12)		0.26** (0.13)
Doctorate		0.40*** (0.12)		0.39*** (0.13)
Income 25,000-49,999		−0.08 (0.07)		−0.08 (0.07)
Income 50,000-74,999		0.05 (0.07)		0.04 (0.07)
Income 75,000-99,999		0.01 (0.07)		0.01 (0.07)
Income 100,000-124,999		−0.13 (0.09)		−0.13 (0.09)
Income 125,000-149,999		0.13 (0.15)		0.13 (0.15)
Income 150,000-174,999		−0.05 (0.07)		−0.04 (0.08)
Income 200,000 and up		0.05 (0.09)		0.05 (0.09)
Treatment (Yes) x Gender (F)			0.07 (0.07)	0.04 (0.08)
Constant	0.40*** (0.03)	0.08 (0.15)	0.41*** (0.04)	0.09 (0.15)
Observations	140	140	140	140
R^2	0.04	0.25	0.05	0.25
Adjusted R^2	0.03	0.11	0.02	0.11
Note:				*p<0.1; **p<0.05; ***p<0.01

Conclusion

After carrying out our experiment and analyzing the results, we can conclude that our treatment had a positive effect on test scores, but there was no evidence of a heterogeneous treatment effect by gender. We also saw evidence of differential attrition, which we theorize as perhaps treatment subjects actually found our treatment to be motivating and it encouraged them to take and to complete the test. However, due to the way we structured our survey, we could not determine what was driving the higher attrition in the control group. Previous research surrounding stereotype threat identified a negative treatment effect for females, so we recommend taking the results from our study and treating it as a hypothesis for future experiments.

Citations

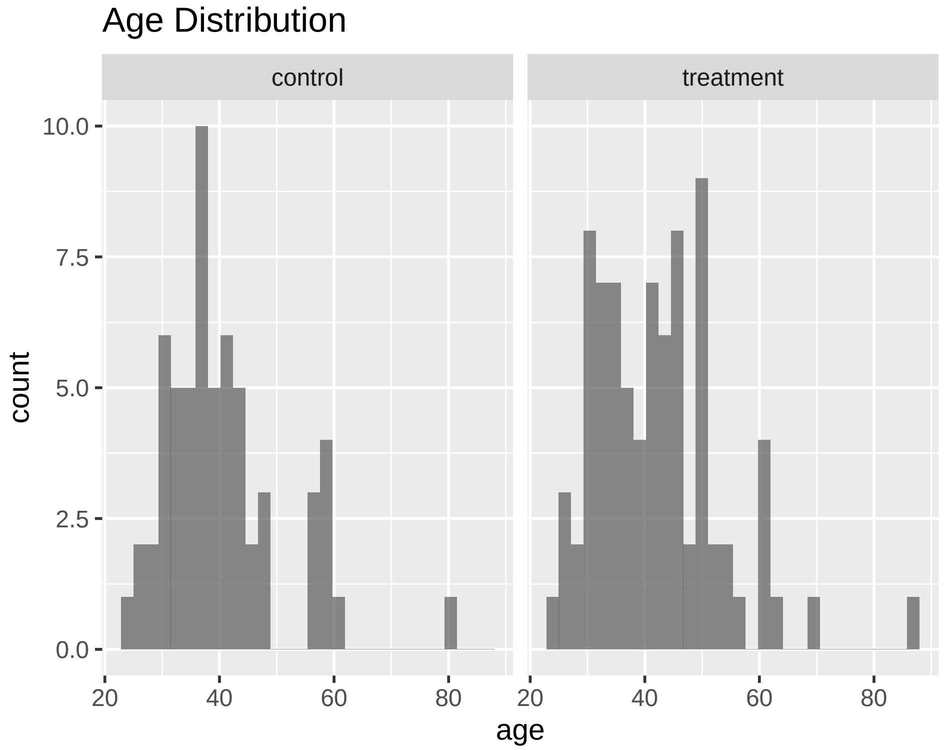
Steven J. Spencer, Claude M. Steele, Diane M. Quinn (1999). Stereotype Threat and Women's Math Performance, Journal of Experimental Social Psychology, Volume 35, Issue 1,1999 https://doi.org/10.1006/jesp.1998.1373.

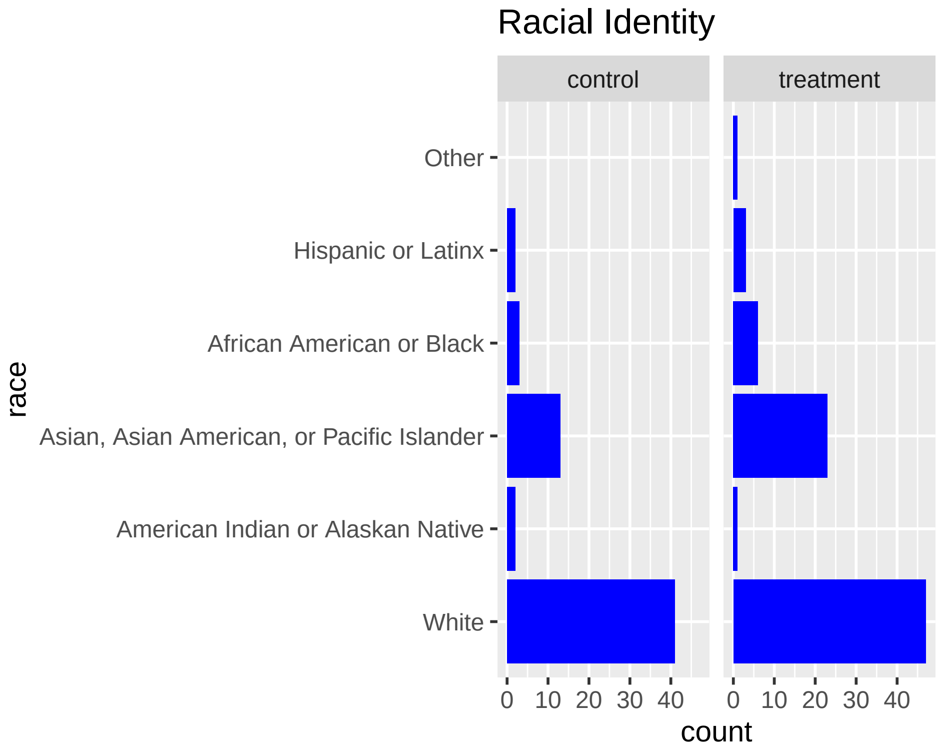
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology, 69(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797.

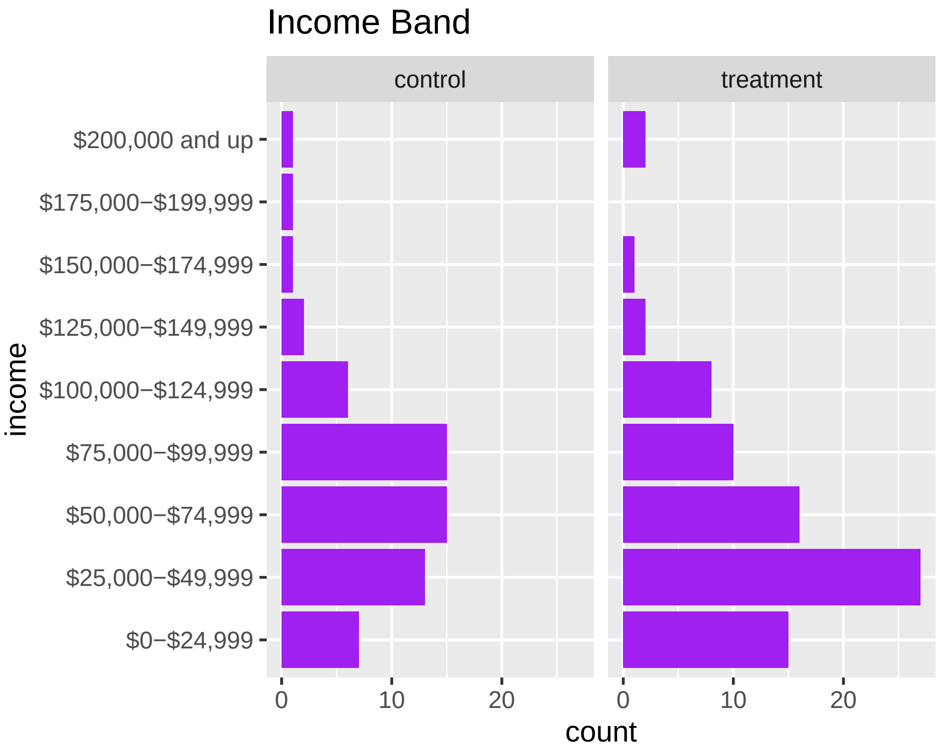
Appendix A

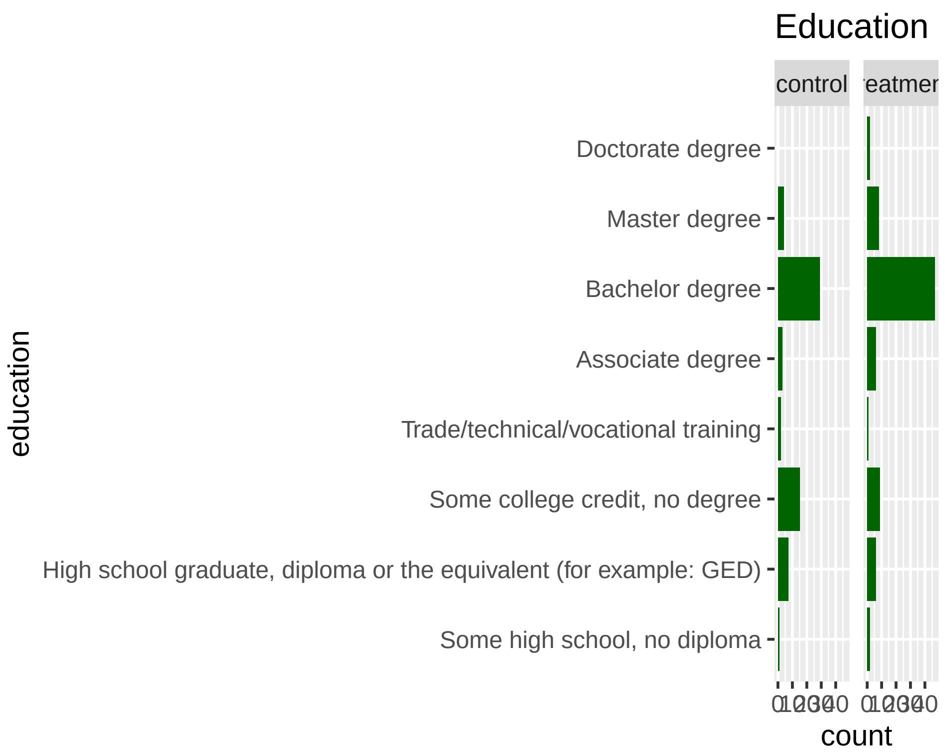
Below are the tables and charts we used to assess whether or not we successfully randomized.

```
##    treatment gender  N
## 1:         1   Male 49
## 2:         0 Female 20
## 3:         1 Female 32
## 4:         0   Male 41
```




Racial Identity
control
treatment
race
Other
Hispanic or Latinx
African American or Black
Asian, Asian American, or Pacific Islander
American Indian or Alaskan Native
White
0
10
20
30
40
0
10
20
30
40
count


Income Band
control
treatment
income
$200,000 and up
$175,000−$199,999
$150,000−$174,999
$125,000−$149,999
$100,000−$124,999
$75,000−$99,999
$50,000−$74,999
$25,000−$49,999
$0−$24,999
0
10
20
0
10
20
count



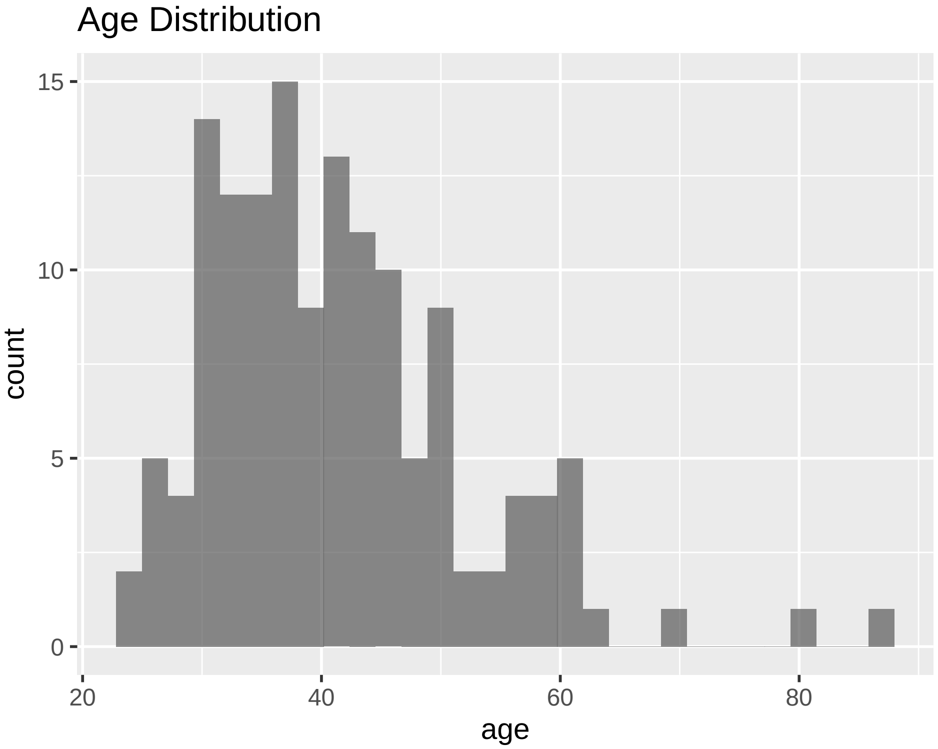
```
## # A tibble: 6 x 5
## # Groups:   gender, race, education [4]
##   gender race  education                                            income count
##   <fct>  <fct> <fct>                                                <fct>  <int>
## 1 Male   White Bachelor degree                                      $25,0~     7
## 2 Female White Bachelor degree                                      $75,0~     7
## 3 Male   White High school graduate, diploma or the equivalent (fo~ $25,0~     6
## 4 Male   White Some college credit, no degree                       $25,0~     6
## 5 Male   White Bachelor degree                                      $0-$2~     6
## 6 Female White Bachelor degree                                      $50,0~     5
```

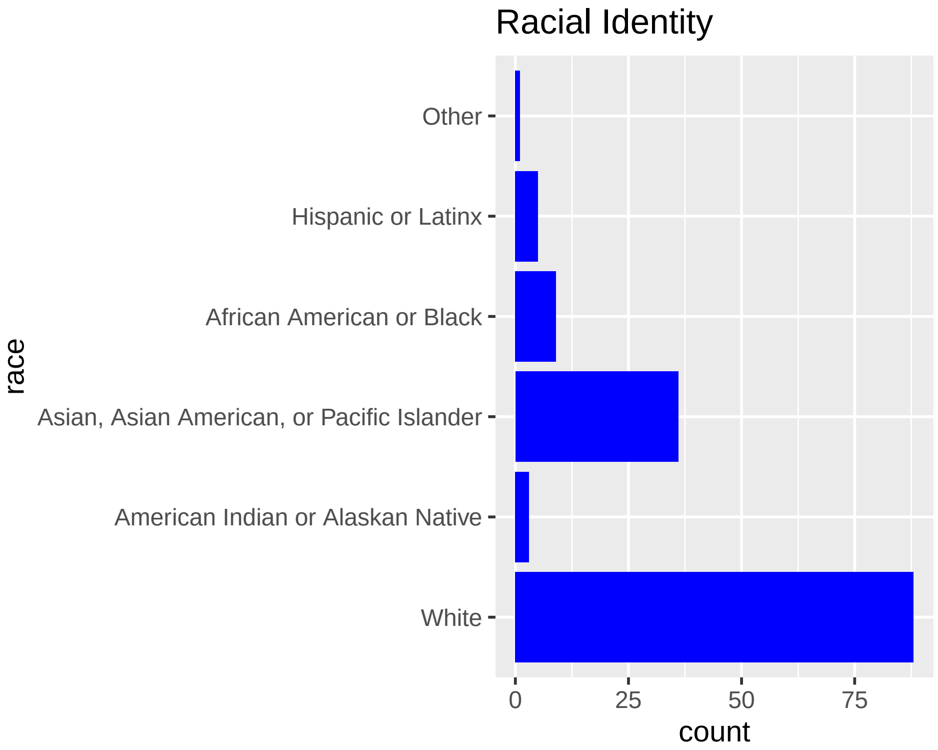
Appendix B

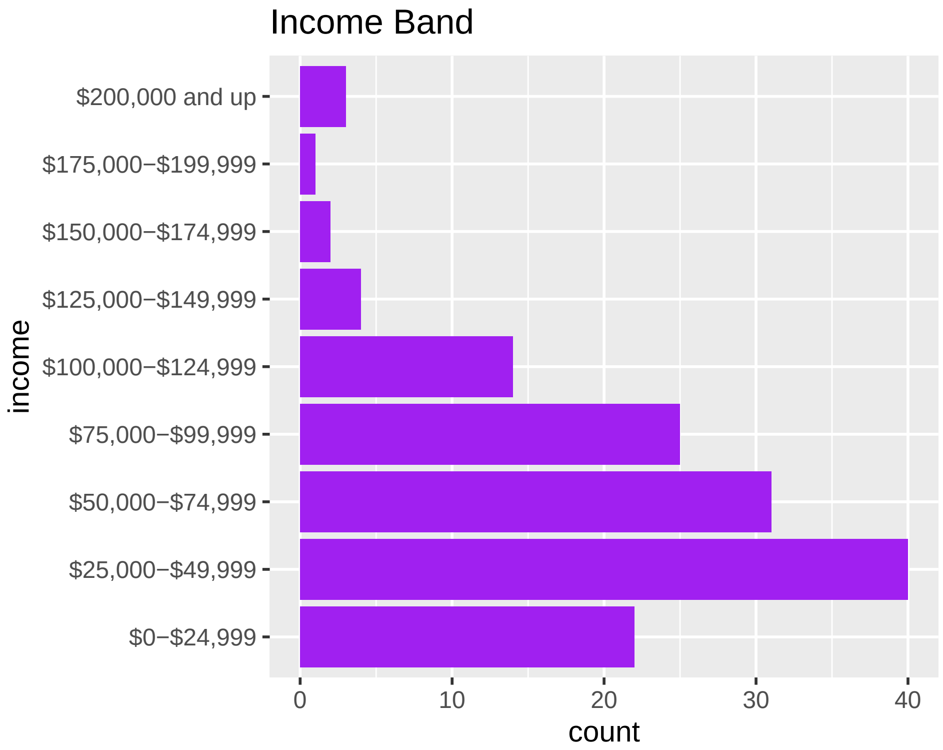
Below are tables and plots that we created as part of our EDA phase to better understand the data.

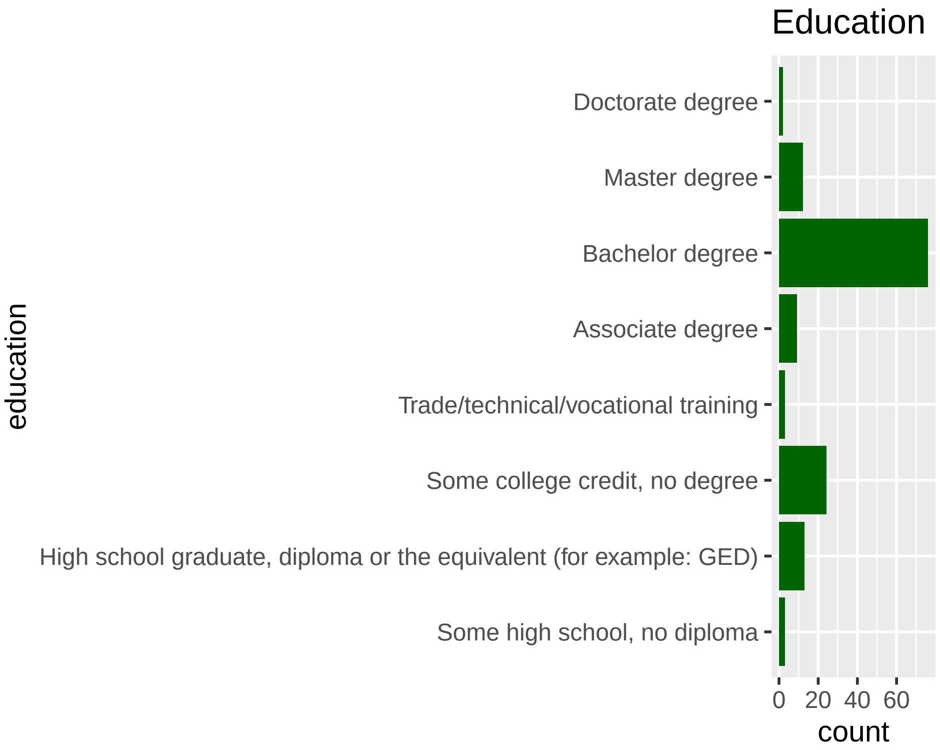
```
##    assignment mean_score
## 1:  treatment  0.4814815
## 2:    control  0.4016393
```

```
##    gender mean_score
## 1:   Male  0.4466667
## 2: Female  0.4480769
```




Racial Identity
Other
Hispanic or Latinx
African American or Black
Asian, Asian American, or Pacific Islander
American Indian or Alaskan Native
White
race
0
25
50
75
count

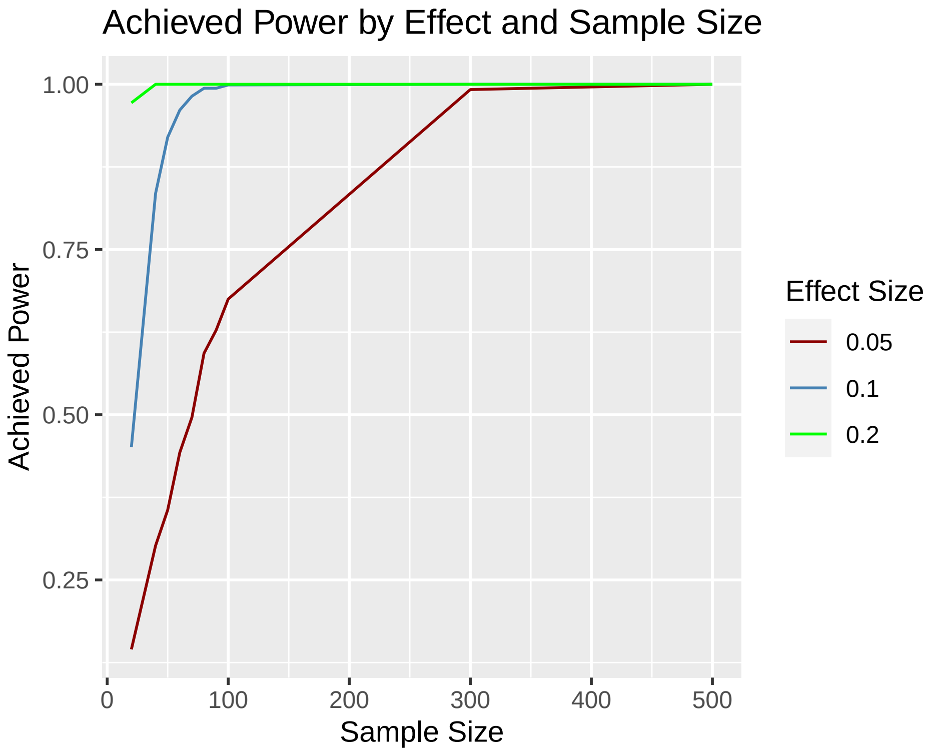

Income Band
$200,000 and up
$175,000−$199,999
$150,000−$174,999
$125,000−$149,999
$100,000−$124,999
$75,000−$99,999
$50,000−$74,999
$25,000−$49,999
$0−$24,999
income
0
10
20
30
40
count


Education
education
Doctorate degree
Master degree
Bachelor degree
Associate degree
Trade/technical/vocational training
Some college credit, no degree
High school graduate, diploma or the equivalent (for example: GED)
Some high school, no diploma
0
20
40
60
count

Appendix C

In this experiment, we are examining the effect of stereotype threat on the test performance of females vs. males. We identified our experiment as potentially having heterogeneous treatment effects, so we will be using a regression model with two main effects (gender and treatment) and an interaction effect. In particular, we believe that our treatment (asking participants for their gender prior to administering a math test) will affect females and males differently. In addition, we will be assessing the significance of the interaction effect using robust standard errors.

From previous studies, we found that the difference in exam scores between males and females in treatment was about 20 percentage points, while there was little to no difference in exam scores between males and females in control. Since these results were from a study where the participants were told that the exam produced differences between females and males, the treatment effect may be higher. For this reason, we are also testing lower treatment effects since we are not including that statement.


Achieved Power by Effect and Sample Size
Achieved Power
Sample Size
Effect Size