Implicit Bias, Gender, and Leadership

Salim Damerdji, Roberto Lievana, Aditya Jhanwar

1. Introduction

In 2015, Time Magazine reported in the US that women were more likely to earn a college degree than men. This trend continues up to this date, the numbers of women earning college degrees keeps rising and now more than ever before women are becoming a larger part of the workforce. However, women are greatly outnumbered by men in leadership positions. In the medical field, 40% of all surgeons and physicians are women, while only 16% are medical school deans. In academia, for the last eight years, women have been responsible for earning the most doctorate degrees, but women account for only 32% of full time professor jobs and only 30% of college presidents or top university representatives. In the financial sector, women account for more than half of the accountants, auditors and financial managers; but women account for only 12% of CFOs at Fortune 500 companies (Boesch et al 2018).

These empirics make clear that, across a wide swath of fields, it is less common for leadership positions to be occupied by women. Part of the barrier to entry for women is related to attitudes and stereotypes about gender and leadership. In particular, there are stereotypes that males make for more natural leaders than women.

Our experiment aims to test avenues for mitigating these implicit biases. In particular, we test two types of arguments about diversity to see their effect people's short-term implicit bias against women occupying leadership positions. To measure implicit bias, we use an abbreviated Implicit Association Test (IAT) modelled after Harvard's IAT. The IAT offers an indirect measure the strength of a person's automatic association of two concepts. We use an IAT to measure how strongly a subject associates one gender with leadership.

Our results are evidence that the two arguments we examine have no marginal effect on short-term implicit bias.

2. Methodology

For our study, we recruited 48 volunteers, all of whom were STEM undergraduate or graduate students at UC Berkeley. We asked participants in the treatment groups to read an argument about female leadership. Afterwards, each participant took an IAT.

Treatments

We studied whether we could alter a person’s implicit bias against female leadership by having participants read one of two arguments about female leadership.

Here is the pro-diversity argument: 'The University of California-Davis reported that the top 25 California companies with the highest percentage of women executives and board members saw a 74 percent higher return on assets and equity than the broader set of companies surveyed. This included companies such as William-Sonoma, Yahoo!, and Wells Fargo.' (Lemke 2019).

Here is the anti-stereotype argument: 'We often think of leaders as dominant and ambitious—as embodying qualities that closely match the stereotype of men. On the other hand, the traits that make up the feminine stereotype (e.g., friendliness and sensitivity) are seen as less vital to leadership. These stereotypes result in women being evaluated less positively than men for leadership positions.' (Prime 2005).

(We acknowledge our way of referring to these arguments is imperfect insofar as neither argument explicitly states that diversity is good or that stereotypes are bad. We use these descriptions only for shorthand convenience.)

In our experiment, we studied two factors that correspond to the two above arguments. Factor A had two levels: the computer program either showed the subject the pro-diversity argument, or it did not. Factor B parallels factor A, but for the anti-stereotype argument. In other words, each subject received one of four treatment combinations: *prior to taking the IAT, the subject either read nothing, read only the pro-diversity argument, read only the anti-stereotype argument, or read both*.

We chose these two arguments to study because they are archetypes of two different types of arguments often made about gender diversity in leadership positions. The pro-diversity argument is one often made in the business world; the argument sells diversity as a route to organizational success. The second argument, in contrast, is more often made in the context of acknowledging the injustice of giving women the short stick of this gender stereotype. Due to how common both of these types of arguments are, we chose to study them.

Units

The units of our experiment were students. Each student only took the IAT once because repeated measurements would threaten our ability to evaluate unconscious biases. That is because, upon taking a complete IAT exam, a student would probably figure out the purpose and structure of the IAT exam, and this knowledge would alter their results on future exams. Thus, we did not perform before-and-after measurements, or any other form of repeated measurement.

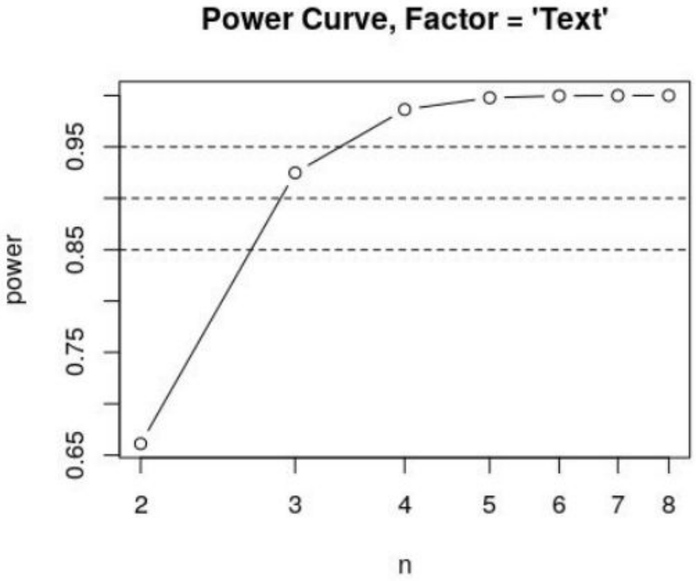
Blocking

Of our 48 volunteers, we had 24 self-identified female participants and 24 self-identified male participants. We blocked by gender because it wouldn’t be surprising if women internalized stereotypes about women differently than men internalized those stereotypes. After all, upbringing and socialization during youth is quite related to one’s gender. By blocking along gender, we planned to convert this unplanned systematic variability into planned variability, reducing the noise in our experiment and increasing our power. In the results section, we show this blocking scheme turns out to be extremely useful.

Power

To increase the power of our experiment, we wanted multiple replications of each of the eight treatment-block combinations. Thus, we opted for a GCB[2] design instead of a CB[2] design, where each treatment-block combination contains only a single sample. We wanted to maximize the power of our experiment through determining a suitable number of observations per treatment. We were aiming for the experiment to have a power of at least 0.8 as per the suggestion of academic research as well as developing a robust experiment all the while choosing a logistically feasible number of individuals to run as subjects given our time constraints.

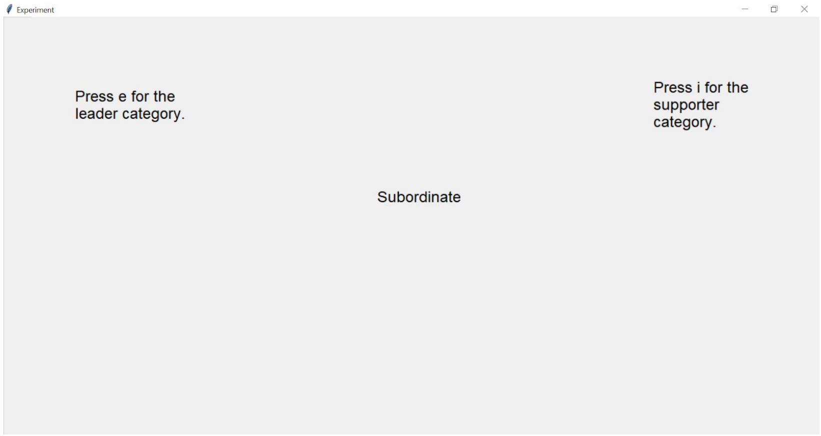
We estimated effects of the different factors after running a few observations within each treatment. We used the carefully obtained data to estimate treatment effects and all other derived estimates and calculations. We assumed a 5 percent significance level, equal variance of errors for both genders, and equal differential effects in the response of approximately 0.3 for the two levels of both the factors (which we decided on based upon calculated results from our data as well as hypothesizing about what the expected effects would be for the full-scale experiment). The plot below shows a single factor's resulting power curve, as a function of the number of individuals per treatment / block combination.



Measurement

To measure the response, we gave participants an abbreviated IAT that consisted of five stages. In stage one, subjects press 'e' for male names and one key for female names. In stage two, it is similar, but for words that are categorized as related to leaders or supporters. In stage three, subjects press 'e' for male names or leader-related words, and 'i' for female names or supporter-related words. In stage four, subjects press 'e' for supporter-related words, and 'i' for leader-related words. In stage five, subjects press 'e' for male or supporter-related words, and 'i' for leader-related words or female names.

Our computer program tracks the delay between when a particular stimulus (e.g. the name 'Joe') is displayed on the center of the screen, and when the subject presses the appropriate key on the keyboard (e.g. 'e'). A screenshot of this setup is displayed in figure 2. This delay is measured in seconds and is rounded to the 5th decimal place. Thus, our measurements are accurate up to the the tenth of a millisecond.



Press e for the leader category.

Press i for the supporter category.

Subordinate

Figure 2

Only the reaction times for stages 3 and 5 are relevant for the response. For those stages, we remove all reaction times longer than 10 seconds, as is standardly practiced in the IAT literature (Flint et al. 2013). These extremely slow responses are indications that participants became distracted or needed clarifications on the instructions. These outliers account for 0.72% of measured reaction times for stage 3 and 5. (That is 14 of 1920 measurements.)

From the remaining reaction times for stage 3 and 5, we subtract the average reaction times stage 3 from that of stage 5. We then divide this difference by the standard deviation for the reaction times across stage 3 and 5. We do this following the lead of experts who argue that “magnitudes of differences between experimental treatment means are often correlated with variability of the data from which the means are computed” (Greenwald et al. 2003). In our dataset, we found a large association of r = .54 between these the magnitude of differences and the variability of the data. This adjustment makes sense since we want our IAT score to be valid and reflect the signal of implicit bias, not the noise of variability in response times.

In short, our response of interest, the IAT score, is the mean reaction time for stage 3 minus the mean reaction time for stage 5, divided by the standard deviation of reaction times for those two stages. The basic intuition is that if subjects have quicker reaction times for stage 3 than for stage 5, then it’s easier for the subject to mentally associate male names and leadership stimuli than female names and leadership stimuli. A large, positive IAT score indicates greater levels of implicit bias.

Protocol

We performed complete randomization within each block. To assign treatments to subjects, we generated two lists where each of the four treatment combinations were listed four times. Both of these lists were randomly rearranged in R. We used the first list to assign treatment combinations to males as they volunteered to participate, and we used the second list for females.

As subjects arrived for the experiment, we used R to randomly assign them to one of the available computers. On that computer, we opened the IAT computer program we coded. Then we input the correct factor combination for that subject, without allowing the subject to see the treatments being assigned. We instructed subjects that all instructions were delivered by the computer program.

The computer program displayed the arguments, if any, that the subject was asked to read. If the subject was assigned to read both arguments, the program, in effect, flipped a coin to decide which to display first. Furthermore, for each stage of the IAT, we randomly order the possible stimuli to display on the screen. We perform these randomizations as insurance against unknown systematic bias, not because we suspected they *would* bias our results.

Validity

Like many other analyses of complicated human responses, it's difficult to study implicit bias with a single measure. There are limitations to the response we have chosen.

First, IAT scores are merely proxies for implicit bias. Virtually all measures of human attitudes are based on indirect measures - including self-reporting or behavior - and our proxy is likewise indirect. Thus, it is a fair question whether high IAT scores are relevant indications of implicit bias.

A 2009 meta-study by Greenwald et al. found that, across 122 studies and 14,900 subjects, IAT scores positively correlated with "a wide range of criterion measures, from interracial friendliness and impression formation to anxious and shy behaviors, consumer choices, and voting," with an average pearson coefficient of .274. These findings held up across a plethora of different types of IATs, including IATs designed to test implicit gender bias.

A correlation coefficient of .274 is not high, but neither are correlation coefficients for other proxies for bias, including self-reporting (Greenwald et al. 2009). Moreover, IAT scores are more reliable than self-reporting for socially sensitive topics (Greenwald et al. 2009). That's because IAT scores can capture implicit biases that subjects may be unconscious of or unwilling to report on a survey.

Second, our results are limited in that they only measure the *marginal* effect of hearing the pro-diversity or anti-stereotype argument one time. It's entirely possible - indeed, likely - that our participants, UC Berkeley students, have heard arguments similar to the pro-diversity or anti-stereotype argument before. Thus, we are not able to study the effect of hearing these arguments repeatedly or, even, for the very first time.

Nevertheless, we believe it's still useful to study the marginal effect of these arguments. Given that arguments similar to the ones we study are fairly common, it's worth learning whether there are gains in gender diverse leadership to be made by repeating these arguments. Moreover, if we suppose that an argument's efficacy fades with time, then the marginal effect of these arguments can inform us about the argument's total effect.

Replicability

The response in question is fairly reliable. Although there is noise in a person's reaction times, this noise is reduced by taking the average of repeated measurements. So even if a person does not react with the exact same speed across repeated trials, their average reaction time should be approximately similar when they take the the test under the same conditions.

Helpfully, there is no measurement error because a computer can measure response times both accurately and precisely without the influence of human error. This helps improve the replicability of our response, reducing noise in our experiment.

Other Words of Caution

First, we only measure short-term impacts on implicit bias. We had participants complete the IAT immediately after treatments were applied. Thus, our response could only measure impacts on implicit bias in the short-term. To many people, it is of greater interest whether these arguments impact implicit bias in the long-term. Due to practical limitations, we were not able to follow up with participants to study the effect of the treatment. Nevertheless, our results are still of interest in that we should not expect long-term effects unless there are short-term effects.

Second, our experiment does not prove there *exists* implicit bias against female leadership. For all we can show, the IAT scores are inflated because participants struggled to adapt to using a new key to press for categorizing gendered names. To rectify this limitation, we would had to have randomized the stages of our IAT test such that stages 2 and 3 could be ordered after stages 4 and 5. Nevertheless, our experiment tests whether our treatments affect the *relative* levels of implicit bias, though we make no claims about the absolute levels of that bias.

Third, UC Berkeley STEM students are not representative of the general population in America. While our motivation is related to larger trends in America, we of course also care about how Berkeley STEM students respond to the arguments we have studied. Furthermore, our experiment can help inform a larger project whereby other researchers study other subsets of the general population for the same phenomena, and a literature review could confirm or deny whether our findings hold generally.

Fourth, the arguments we study are merely archetypes of the types of arguments we are interested in. However, our results cannot be quickly generalized to other arguments that are substantially different. For example, if we find a significant effect for the prodiversity argument, it may be due to the tone, the use of empirics, or the framing of the argument. Our study is not able to exactly discern these differences. Nevertheless, it is still a worthwhile to see whether this argument has a significant effect on implicit bias because it serves as an indication about the prospects of other similar arguments.

Data Table

In the data table below, subject_id is a unique identifier for each participant, time refers to when the test was started, 'cp' denotes which of three computers the subject took the IAT on, 'prodiversity' is a column of booleans that indicate whether the subject read the prodiversity argument, 'anti-stereotype' is likewise a column of booleans that indicate whether the subject read the anti-stereotype argument, and 'iat' denotes the subject's IAT score.

subject_id	time	gender	cp	prodiversity	antistereotype	iat_d
1	10:31:06	Male	2	FALSE	TRUE	0.103
2	10:34:22	Female	2	FALSE	TRUE	-0.166
3	10:45:03	Female	2	TRUE	FALSE	0.308
4	10:48:56	Female	2	TRUE	TRUE	0.672
5	10:52:27	Male	1	FALSE	FALSE	0.16
6	11:42:21	Male	3	TRUE	TRUE	0.049
7	12:09:09	Female	3	TRUE	TRUE	-0.545
8	12:15:50	Male	3	FALSE	TRUE	0.641
9	12:17:19	Male	1	TRUE	TRUE	0.798
10	12:18:55	Female	3	FALSE	FALSE	0.066
11	12:21:36	Female	1	TRUE	FALSE	-0.114
12	12:21:59	Male	3	FALSE	FALSE	0.431

13	12:24:52	Female	3	TRUE	TRUE	-0.128
14	12:26:50	Male	1	TRUE	TRUE	-0.194
15	12:28:24	Female	3	TRUE	TRUE	0.618
16	12:30:56	Male	3	FALSE	TRUE	0.003
17	12:32:35	Female	2	FALSE	TRUE	0.051
18	12:32:44	Male	1	FALSE	FALSE	-0.433
19	12:33:00	Male	3	FALSE	TRUE	0.571
20	12:37:44	Male	3	FALSE	FALSE	0.684
21	12:39:01	Male	1	FALSE	FALSE	0.169
22	12:42:25	Female	3	TRUE	FALSE	-0.429
23	12:43:46	Male	1	FALSE	TRUE	-0.209
24	12:46:01	Female	3	TRUE	FALSE	0.618
25	12:47:01	Male	1	TRUE	TRUE	0.006
26	12:53:05	Male	2	TRUE	TRUE	-0.216
27	13:04:00	Male	3	TRUE	FALSE	0.637
28	13:08:03	Female	3	TRUE	FALSE	0.022
29	13:08:45	Female	2	FALSE	FALSE	-0.059
30	13:13:36	Female	1	TRUE	TRUE	-0.099
31	13:15:00	Female	2	FALSE	TRUE	-0.662
32	13:15:24	Female	3	TRUE	TRUE	-0.59
33	13:16:57	Female	1	FALSE	FALSE	-0.221
34	13:19:44	Male	3	TRUE	FALSE	-0.355
35	13:22:47	Female	1	FALSE	TRUE	0.058
36	13:34:05	Female	3	TRUE	FALSE	-0.281
37	13:43:21	Male	1	FALSE	TRUE	0.356

38	15:06:18	Female	2	FALSE	TRUE	-0.148
39	15:10:15	Female	2	FALSE	FALSE	-0.451
40	15:16:58	Female	2	FALSE	TRUE	-0.023
41	15:19:44	Female	2	FALSE	FALSE	-0.47
42	15:26:03	Female	2	FALSE	FALSE	0.943
43	18:44:15	Male	3	TRUE	FALSE	0.13
44	22:25:03	Male	3	TRUE	FALSE	0.354
45	22:28:54	Male	3	TRUE	FALSE	0.115
46	22:32:27	Male	3	FALSE	FALSE	0.78
47	22:33:35	Male	1	TRUE	FALSE	0.475
48	22:37:33	Male	3	TRUE	TRUE	0.907

Figure 3

3. Results

Checking Model Assumptions

Before conducting any analysis or exploring the data we first wanted to determine if the assumptions of a parametric model are met. Several assumptions for such a model we wanted to test were normally distributed and homoskedastic errors as well as constant effects and additivity.

We created two interaction plots to better understand if an additive or interaction model would best suit our observed data. Figure 4 is a 2-way interaction plot that highlights the interaction effects of the factors of interest, whereas figure 5 is a 3-way interaction plot that accounts for the blocks of male and female genders. For added clarity in analysis, we included

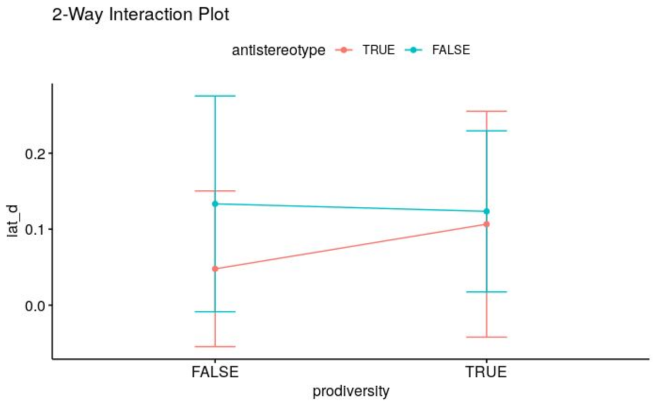


Figure 4

standard error bars to account for variability in the data to better gauge the plot.

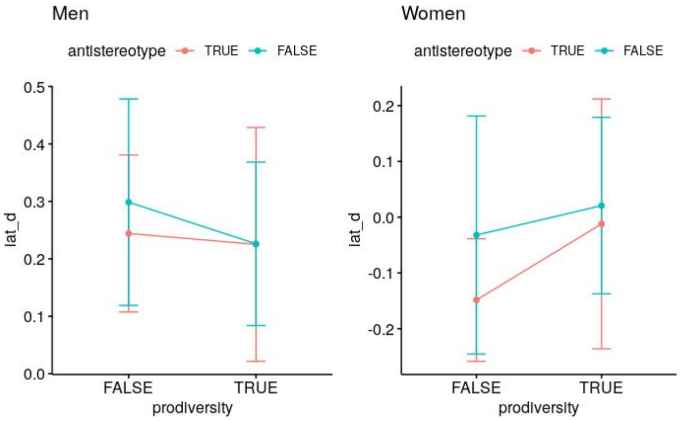


Figure 5

As we can see in both the 2-way and 3-way interaction plots, the ‘lines’ do not appear to be parallel. This was also the case for additional 2-way interaction models that were produced as well. However, when accounting for the variability of the data as displayed by the standard error bars we noticed that we did not have enough data to accurately determine whether there is in fact some sort of interaction prevalent. Hence, after careful analysis and observation of the plots we reasoned that we would assume an additive model as the noise in the ‘lines’ could perhaps be explained by the variability itself rather than through an interaction term.

We implemented a fitted versus residual plot as it serves as a robust method in examining several model assumptions such as constant variance and zero-mean errors.

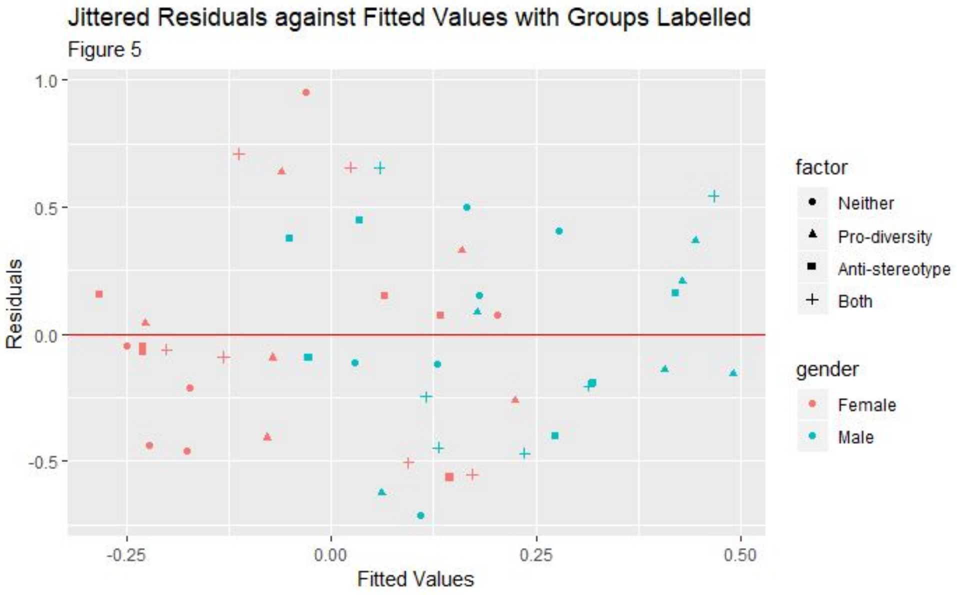

Jittered Residuals against Fitted Values with Groups Labelled

Figure 5

From observing the fitted versus residual plot that was produced given our data we found the errors to be fairly homoskedastic, which argued that the aforementioned assumptions were in fact met. In our judgment, this held true within blocks as well; although variance decreases among large fitted values for males, this coincides with fewer observations, so it is not strong evidence that any assumptions of ANOVA are violated.

We also wanted to confirm that the errors are normally distributed as per another important assumption of the model and ANOVA analysis. We produced a Q-Qplot as it would best indicate a violation in the normality assumption.

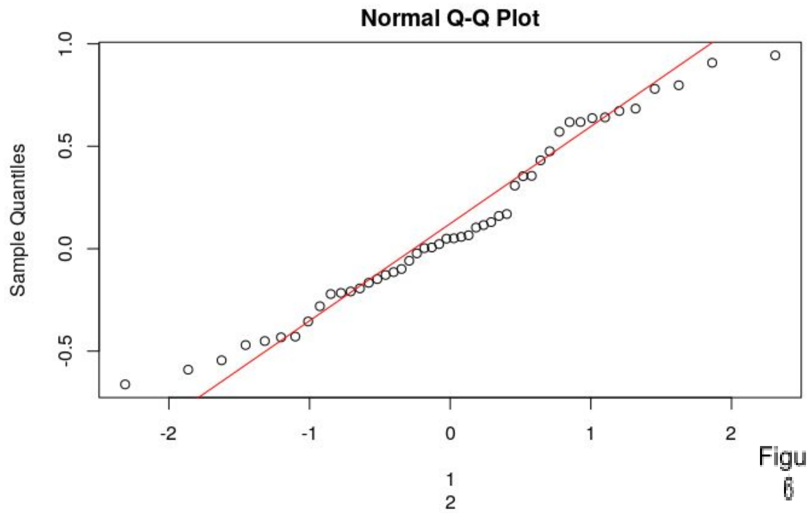


Figure 7
6

The results produced showed that the errors are distributed fairly normally albeit with some noise present. However, there does not seem to be any outliers and seems like a satisfactory match for the assumption. We did try using transformations on the data to form much more stringently normally distributed data, which would lead to that much more of an effective and representative model, but ultimately were unsuccessful of finding any such transformation and we felt the results produced were satisfactory.

Independence Assumption

In addition to analyzing the standard model assumptions we also took a look at other aspects of the data for sources of bias and variability. Our experiment took place mainly during 10 AM - 2 PM, which was the assigned time for running our experiment. Since we were unable to recruit all subjects necessary for the experiment during the time, we gathered the remaining subjects necessary afterwards outside of lab hours. We were concerned of possibility of the times subjects performed the experiment having some effect with the response times. We also were concerned with which computer the IAT was taken on. Fortunately, neither plot suggests we violate the independence of errors assumption since the residuals all look quite similar.

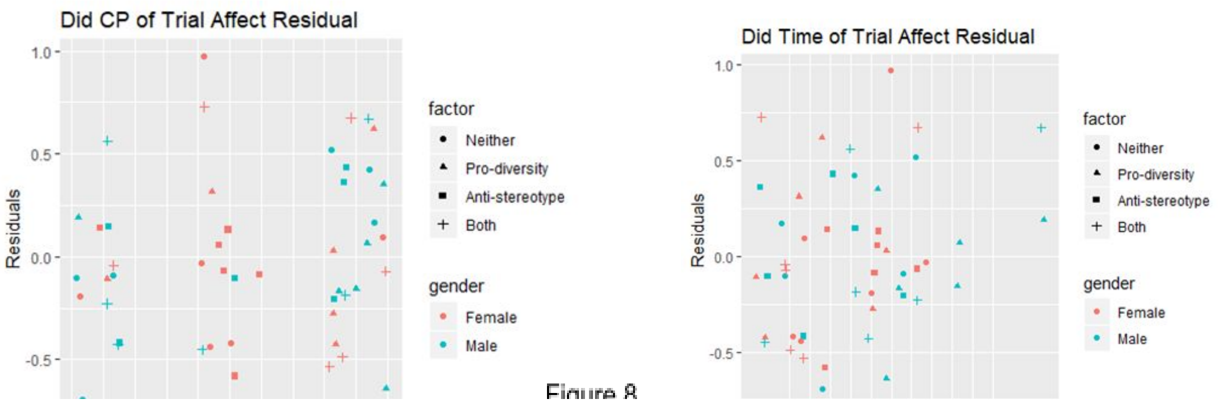


Figure 8

Parallel Dot Chart

As a means of viewing the entire data in a concise method, we produced a parallel dot chart. We observe in the figure below that subjects who identified as female seem to have higher response times in general compared to those who were male. This reaffirms the presence of variability in the responses for the different genders and that blocking was in fact an appropriate choice for our design of the experiment.

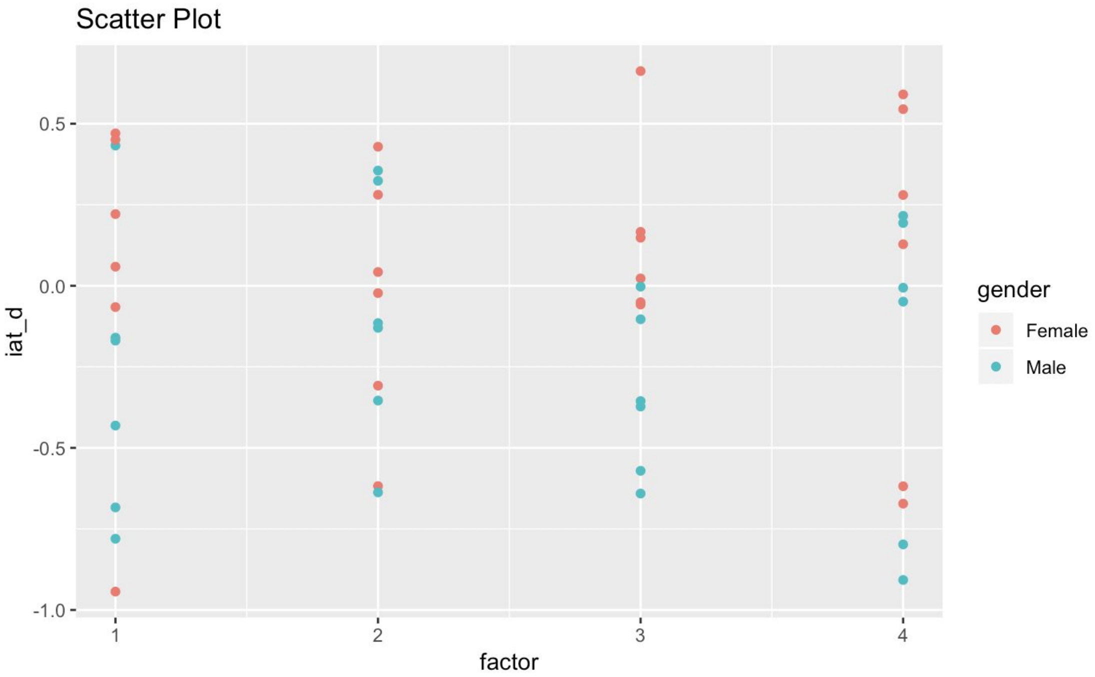


Figure 9

ANOVA

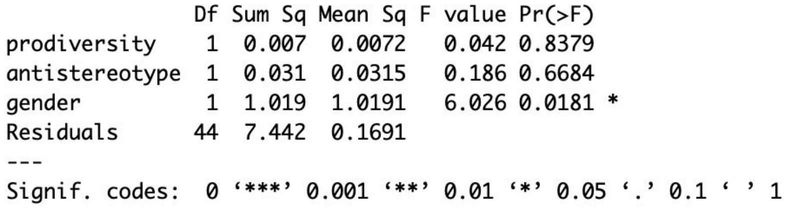
Our model is

$$y_{ijkl} = \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijkl}$$

Where i = 1, 2 and α_i is the Factor A “pro-diversity”, j = 1, 2 and β_j is the Blocking Factor “gender” , k = 1, 2 and γ_k represents Factor B “anti-stereotype”; and *l* = 1 …. 48 and ε_{ijkl} are the error terms. We assume the ε_{ijkl} ~ N (0, σ^2)

In addition we considered gender to be a fixed effect rather than a random effect. Gender is constant across individuals over time. To make it clear, any effects of being a male will not change over time. This means that gender has a fixed effect.

Table 1

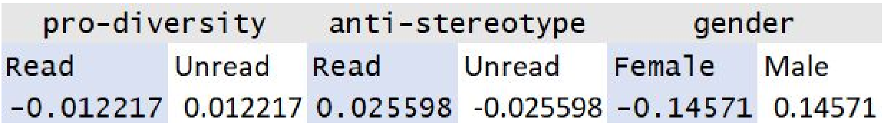


```
                Df Sum Sq Mean Sq F value Pr(>F)  
prodiversity     1  0.007  0.0072   0.042 0.8379  
antistereotype   1  0.031  0.0315   0.186 0.6684  
gender           1  1.019  1.0191   6.026 0.0181 *
Residuals       44  7.442  0.1691                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The only significant effect in our analysis corresponds to the blocking factor “gender” [F(1, 44) = 6.026, P<0.05] from Table 1. Most of the variance that was previously attributed to residuals has now been partitioned to the block effect. This was going to help us explain the differences between the different factors more profoundly, however, the two factors of interest showed minimal effect [F(1, 44) = 0.0072, P>0.05] and [F(1, 44) = 0.0315 , P>0.05] from Table 1.

Here is a table of estimated effects:

Table 2



pro-diversity		anti-stereotype		gender	
Read	Unread	Read	Unread	Female	Male
-0.012217	0.012217	0.025598	-0.025598	-0.14571	0.14571

Effects of Blocking

The blocking strategy in the experiment proved to be successful. Our suspicion, that baseline rates of implicit bias vary by gender was correct. The large variation between gender/blocks is clear with a F-ratio [F(1, 44) = 6.026, P<0.05] from our ANOVA analysis in Table 1. Blocking by gender enabled us to convert this unplanned systematic variability into planned variability which reduced the noise in our experiment and increased the power of our experiment. Any cause due to a difference in sex is taken into consideration by using gender as a blocking factor.

Contrast

Let’s examine the contrast for whether there was a difference in response between the control group and the treatment groups. This corresponds to the coefficient vector of (-3, -3, 1, 1, 1, 1, 1, 1), where first two entries of the coefficient vector correspond to the two control groups (one per gender), and the last four entries correspond to the remaining treatment / block combinations. Our null hypothesis is that there is no difference between the control groups and treatment groups. Our alternative hypothesis is that there is some difference.

The unbiased estimate of this linear combination is 0.0571. We use MS residuals as an estimate for error, by multiplying it by the square root of $\Sigma w_i/n_{i}\cdot$, which is 3.8333.

Thus, our t-statistic is .0571 / sqrt(.1691*3.8333). This t-statistic has a p-value of 0.9445. Clearly, this is not statistically significant. So, we fail to reject the null that there is no difference between the control and treatment groups.

Results and Ideas For Future Research

In short, gender appears to be predictive of IAT scores, but the arguments displayed did not have an effect. This is an interesting finding that we found surprising. It would be natural to assume that these arguments were suggestive enough to reduce implicit bias. Perhaps implicit bias is resilient, or these arguments were unconvincing, or some other possibility could explain this result.

Due to this uncertainty, we believe this was merely a starting point for the many possibilities of these sorts experiment. An issue that was touched on earlier is that our experiment was limited to STEM students at UC Berkeley, which does not necessarily generalize well to the standard characteristics of males and females in the larger scope of the population. An adaption of this experiment in which one would source subjects more openly and not within a single cluster could lead to much more robust results. In addition, our main factors only accounted for different types of text whereas other forms of media could have different implications and effects on the response. With further resources and newfound experience in designing and analyzing an experiment, we suspect that we or even other individuals with more domain knowledge on the subject of the experiment could further expand on the what the motivation behind the experiment is.

References

Dasgupta, Nilanjana, and Anthony G. Greenwald. “On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals.” *Journal of Personality and Social Psychology*, vol. 81, no. 5, 2001, pp. 800–814., doi:10.1037//0022-3514.81.5.800.

Flint, Stuart W., et al. “Counter-Conditioning as an Intervention to Modify Anti-Fat Attitudes.” *Health Psychology Research*, vol. 1, no. 2, 2013, p. 24., doi:10.4081/hpr.2013.738.

Harvard, Project Implicit. https://implicit.harvard.edu/implicit/selectatest.html

Page-Gould, Elizabeth, and Amanda Sharples. *How to Avoid Picking Up Prejudice from the Media*. Greater Good, 7 Sept. 2016, greatergood.berkeley.edu/article/item/how_to_avoid_picking_up_prejudice_from_media.

Thomas, Breda, et al. *Can Female Role Models Reduce the Gender Gap in Science? Evidence from Classroom Interventions in French High Schools*. Working Paper. 2018.

Lemke, Tim. “Do Companies With Female Executives Perform Better?” The Balance, The Balance, 1 Feb. 2019, www.thebalance.com/do-companies-with-female-executives-perform-better-4586443.

Prime, Jeanine. Women “Take Care,” Men “Take Charge.” Edited by Andrea Juncos and Kara Patterson, Catalyst, 2005, pp. 1–38, Women “Take Care,” Men “Take Charge.”

Article on how to compute IAT score:
https://faculty.washington.edu/agg/pdf/GB&N.JPSP.2003.pdf

Article on external validity:
http://www.people.fas.harvard.edu/~banaji/research/publications/articles/2009_Greenwald_JPSP.pdf

Article on Women Leadership Gap
https://www.americanprogress.org/issues/women/reports/2018/11/20/461273/womens-leadership-gap-2/

Article Times

http://time.com/4064665/women-college-degree/

Appendix

~~~
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, comment = "")
library(readr)
library(dplyr)
library(ggpubr)
library(ggplot2)
library(magrittr)
library(gridExtra)
options(contrasts=c("contr.sum","contr.poly"))
~~~

```{r load_data}
data = read_csv("iat_scores.csv", col_types = cols(.default = col_factor(),
time = col_time(), iat_d = col_double()))
```

```{r}
glimpse(data)
```

```{r model}
df = aov(iat_d ~ prodiversity + antistereotype + gender, data = data)
```

```{r fit_resid}
# Fitted vs. Residual
fit_resid = tibble(f = fitted(df), r = residuals(df))

ggplot(data = fit_resid, aes(x=f, y=r)) +
  geom_jitter(width = 0.25) +
  geom_hline(yintercept = 0, color='red') +
  ggtitle("Fitted vd. Residual Plot") +
  xlab("Fitted") +
  ylab("Residual")
```

```{r fit_resid_run_order}
# Run order vs. Residual
run_resid = tibble(run = data$run, res = residuals(df))
ggplot(data = run_resid, aes(x=run, y=res)) +
```

~~~
  geom_point() +
  geom_hline(yintercept = 0, color='red') +
  ggtitle("Run Order vd. Residual Plot") +
  xlab("Fitted") +
  ylab("Residual")
```

```{r qqplot}
# qqplot
qqnorm(x=data$run, y=data$iat_d);
qqline(y=data$iat_d, col='red')
```

```{r interaction_plot_2way}
# Interaction Plot 2 Way
ggline(data, x = "prodiversity", y = "iat_d", color = "antistereotype", add
= c("mean_se")) + ggtitle("2-Way Interaction Plot")
```

```{r interaction_plot_3way}
# Interaction Plot 3 Way
male = data %>% filter(gender == "Male")
female = data %>% filter(gender == "Female")

ip1 = ggline(male, x = "prodiversity", y = "iat_d", color =
"antistereotype", add = c("mean_se")) + ggtitle("Men")
ip2 = ggline(female, x = "prodiversity", y = "iat_d", color =
"antistereotype", add = c("mean_se")) + ggtitle("Women")

grid.arrange(ip1, ip2, ncol=2)
```

```{r}
# Anova table used for analysis
summary(df)
```
~~~