W241 Thursday 4pm

Final Paper

Karen Liang, Song Park, Kenneth Pong

**Gender Bias and Perceived Aggressiveness**

# 

# Abstract

In this experiment, we sought to test for the existence of a bias in perceptions of aggressiveness based on the gender of the individual. To do so, we conducted an experiment delivered via Mechanical Turk and Qualtrics wherein the users read the same story, but with varying protagonist genders in treatment and control (n=832). Measures were taken to control for noncompliance by introducing metrics and sanity checks to filter them out. The results failed to produce sufficient evidence to reject the null hypothesis, and the point estimates indicated in the opposite direction suggested by the research. Possible explanations for this inconclusive result are imperfect administration of the treatment, as well as the respondent pool being drawn from a group unlikely to be representative of the broader population.

# 

# Introduction and Background

There have been many studies showing the underrepresentation of women in the workforce, especially in the technology domain. According to a recent study by McKinsey, for the sixth year in a row, women continued to lose ground at the first step up to manager. For every 100 men promoted to manager, only 85 women were promoted — and this gap was even larger for some women: only 58 Black women and 71 Latinas were promoted. As a result, women remained significantly outnumbered in entry-level management at the beginning of 2020 — they held just 38 percent of manager-level positions, while men held 62 percent [1].

There are many reasons that the underrepresentation of women in the corporate pipeline persists. One aspect is particularly interesting. In an article published by the Harvard Business Review [2], which looked at 200 performance reviews within one company, it revealed gender bias through the lens of performance reviews. The results tallied the number of references to being “too aggressive” in the reviews and, not surprisingly, 76% of the instances were attributed to women, while only 24% of men were identified as having such a communication style.

A similar topic is also covered in a book called *Nevertheless, She Persisted: True Stories of Women Leaders in Tech* written by *Pratima Rao Gluckman.* Gluckman interviewed many women in the workplace for her book. In an article, Gluckman wrote, “after interviewing many women (especially female leaders in the workplace), when a woman speaks her mind and is assertive, she is likely to be labeled as aggressive. So, are these women confronting and attacking their colleagues or are they simply being effective leaders? We find this topic to be profoundly troubling. It is hard to imagine anyone leading effectively without being assertive.[3]”

Does gender bias contribute to the perception of “aggressiveness” of a person? In another word, is there a double standard on “aggressiveness” in the workplace based on gender? We found this a very interesting research question that would potentially help us to get one small step closer to the bigger puzzle of the underrepresentation of women in the workforce.

# Methodology

Experiment Design

To create a controlled experiment with gender as the only variable, all experiment participants are asked to read a short story (~500 words) that describes a workplace conflict. The control group is presented with the male version of the story, with the main character named “Kevin”and referred to as “him” throughout the story. The treatment group is presented with the female version of the story, with the main character named “Kelly” and referred to as “her”. Aside from his/her pronouns and character names, the story is identical across control and treatment groups.

To measure how aggressive participants perceived Kelly/Kevin to be, participants, after reading the survey, are asked to rate the level of aggressiveness of the characters.

Choice of Story

Ideally, given more time and resources, we would be able to use multiple stories with varying levels of aggressiveness and varying scenarios. However, due to practical limitations and the need to produce a sufficient sample size for different demographic groups, we opted to use a single story that we wrote ourselves.

The pick of the story itself was rather important, since it should have some level of controversy yet not have a clear right or wrong response. Additionally, the language and tone of the story needed to be relatively neutral so as not to bias readers’ opinions, and the situation presented in the story needed to be relatable or at least somewhat familiar to the majority of participants.

After much research, we chose to create a compensation discussion scenario (see Appendix A). Compensation discussion is known to be a sensitive subject in the workplace. Depending on the way people approach it, it may sometimes be “uncomfortable” to discuss openly with managers and supervisors, but it is a common topic that individuals with work experience are likely to relate to, and those without work experience are likely to have at least a limited understanding of.

In the final version of the story, the main character’s manager denies the main character’s request for a raise. The main character continues to press his/her case despite the initial denial, creating a scenario that could be viewed as “aggressive.”

Choice of Names

For the main characters in the two different versions of the story, we aimed to select names that are as sentiment-neutral as possible. Despite the balance guaranteed by truly random selection, we wanted to remove at least one known possible source of variance in scores - for instance, names attached to political figures could introduce undue variance into the results. Thus, we took the 100 most common names over the past 100 years as determined by the United States Social Security Administration [4], and searched for their sentiment scores using GloVe pre-trained embeddings [5] trained on the Common Crawl comprising 42 billion tokens. We used text-to-sentiment analysis to extract the sentiment scores, and selected for names closest to a sentiment score of zero. For a partial list of the 20 most sentiment neutral names by gender, see Appendix B.

Through this process, we determined that the most sentiment-neutral male and female names were Kevin (score 0.019) and Kelly (score -0.093), which were used for the name of the main character in control and treatment, respectively. This same exercise yielded a runner-up of Ronald (score -0.025) for the manager character - however, due to the risk of sentiment contamination via conflation with Donald (score -3.829), the third-most neutral male name (Sean - score -0.045) was instead used for the manager character.

Survey Design

At a high-level, the survey can be broken out into 4 sections, with all questions asked in the survey being mandatory:

1. Demographic questions
2. Story
3. Questions about the story
4. Questions about aggressiveness

To begin, we ask a few demographic questions that would be used for blocking, understanding our respondent demographic make-up, as well as for covariates in our regression analysis. The demographic questions asked were gender, age, education level, years of professional work experience, and race.

Following the demographic questions, participants are presented with either the male or female version of the story. After reading the story, the participants were first asked a few questions about the story. We asked participants questions about the story in order to minimize noncompliance. Although total survey time is an indicator for noncompliance, it’s difficult to tell for sure whether or not participants fully read the story before proceeding. Therefore, we asked three recall questions to detect noncompliance and to encourage participants who did not accurately read the story to go back and read the story. An example of one of these recall questions is as below:

1. *What is the raise Kelly/Kevin asked for?* 
   1. *5%*
   2. *10%*
   3. *15%*
   4. *20%*

After answering story recall questions, participants then responded to a few questions on the perceived aggressiveness of the characters in the story. The participants would rate the characters aggressiveness on a scale of 0-10. Although our main outcome variable of interest was the perceived aggressiveness of Kevin/Kelly, we also asked participants about the aggressiveness of Sean (the manager character) as this could be a potential covariate.

Platforms (Mechanical Turk, Qualtrics)

Because our experiment focused on aggressiveness and gender bias in the workplace, we wanted our respondent pool to be mostly individuals with professional work experience. As a result, we opted to recruit participants through Amazon Mechanical Turk (MTurk) instead of using Berkey’s XLAB. MTurk recruitment was limited geographically to those within the U.S.

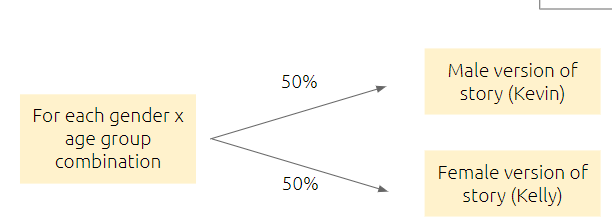
The participants recruited through MTurk were given a link to our survey, which we created through Qualtrics. In addition to the ease of being able to use our Berkeley accounts to access an otherwise paid account level for Qualtrics, Qualtrics also offered easy-to-use methods of blocking and randomization that would allow us to set up the experiment in the way we wanted.

Respondent Recruitment

In terms of sample size, we decided to recruit as many participants as we could by maximizing the budget provided by the MIDS program. This helped us to maximize statistical power, accounting for expected sample attrition from users not engaging in experiment as intended. Although we had initially aimed for 1,000 survey responses, by the end of our one-week recruitment process we secured 832 respondents. While there are many different explanations for this, we believe it’s likely that tasks on MTurk receive the highest engagement immediately after the task is posted, then taper off as new tasks continuously come up. Therefore, the time the task was posted could have influenced the level of initial engagement, resulting in fewer responses than we had initially planned to receive.

Blocking & Random Assignment of Treatment

We decided to use blocking to make sure we received an even number of treatment and control group responses within certain demographic groups. Using the responses to the demographic questions in the survey, we were able to block based on both gender and age group. The survey question for gender had 4 response options (Male, Female, Non-binary/third gender, Prefer not to say) and that for age group had 5 options, resulting in a total of 20 blocks. Within each block, participants were randomly assigned to either the control (male version of the story) or treatment (female version of the story) groups as explained in the visual below.



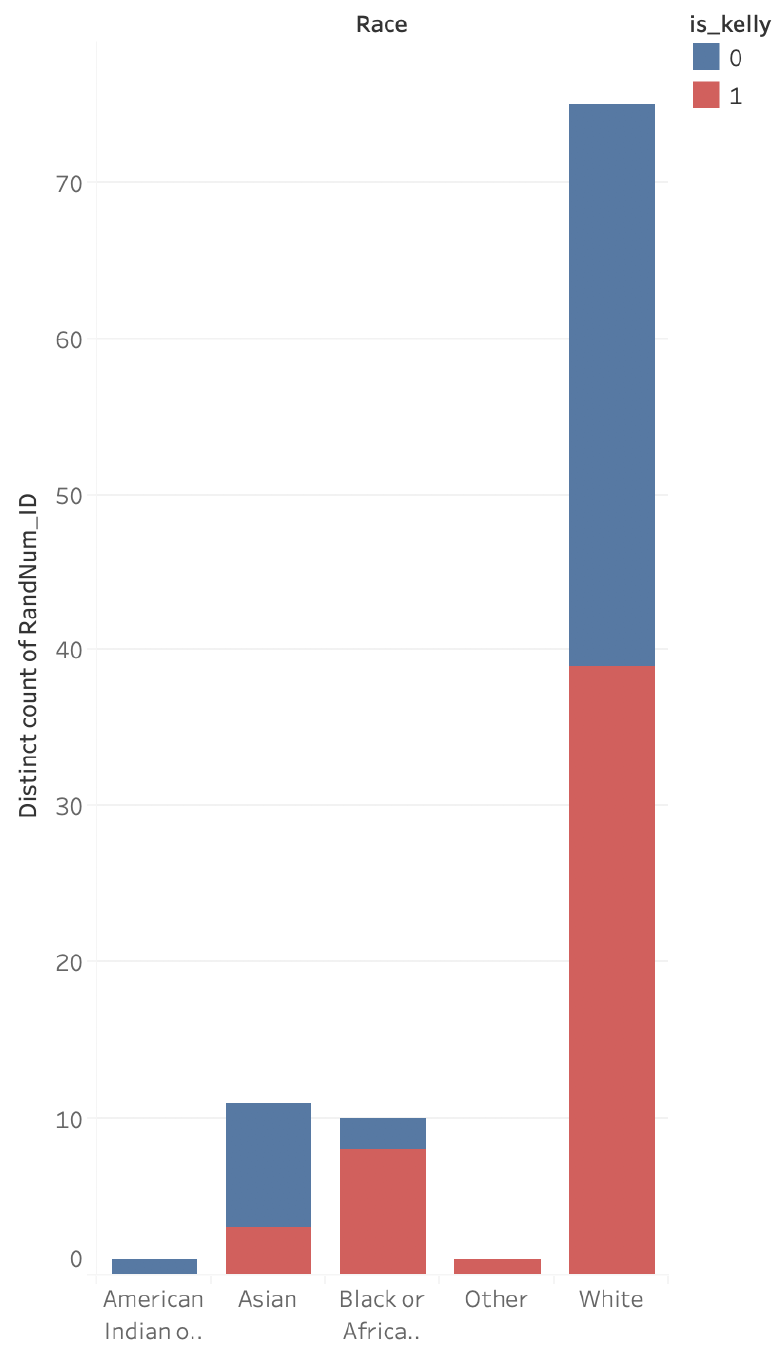
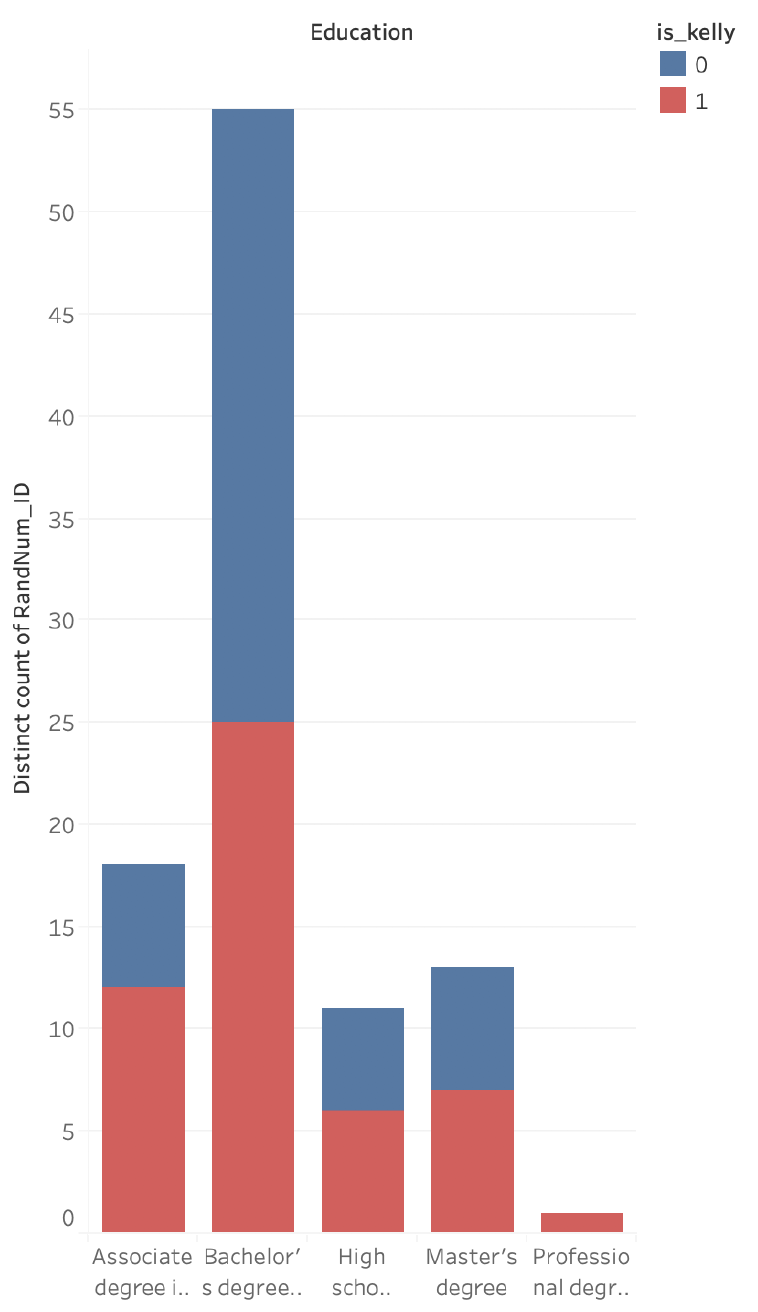
*Figure 1: Treatment and Control Blocking Illustration*

# Pilot Study and Error Detection

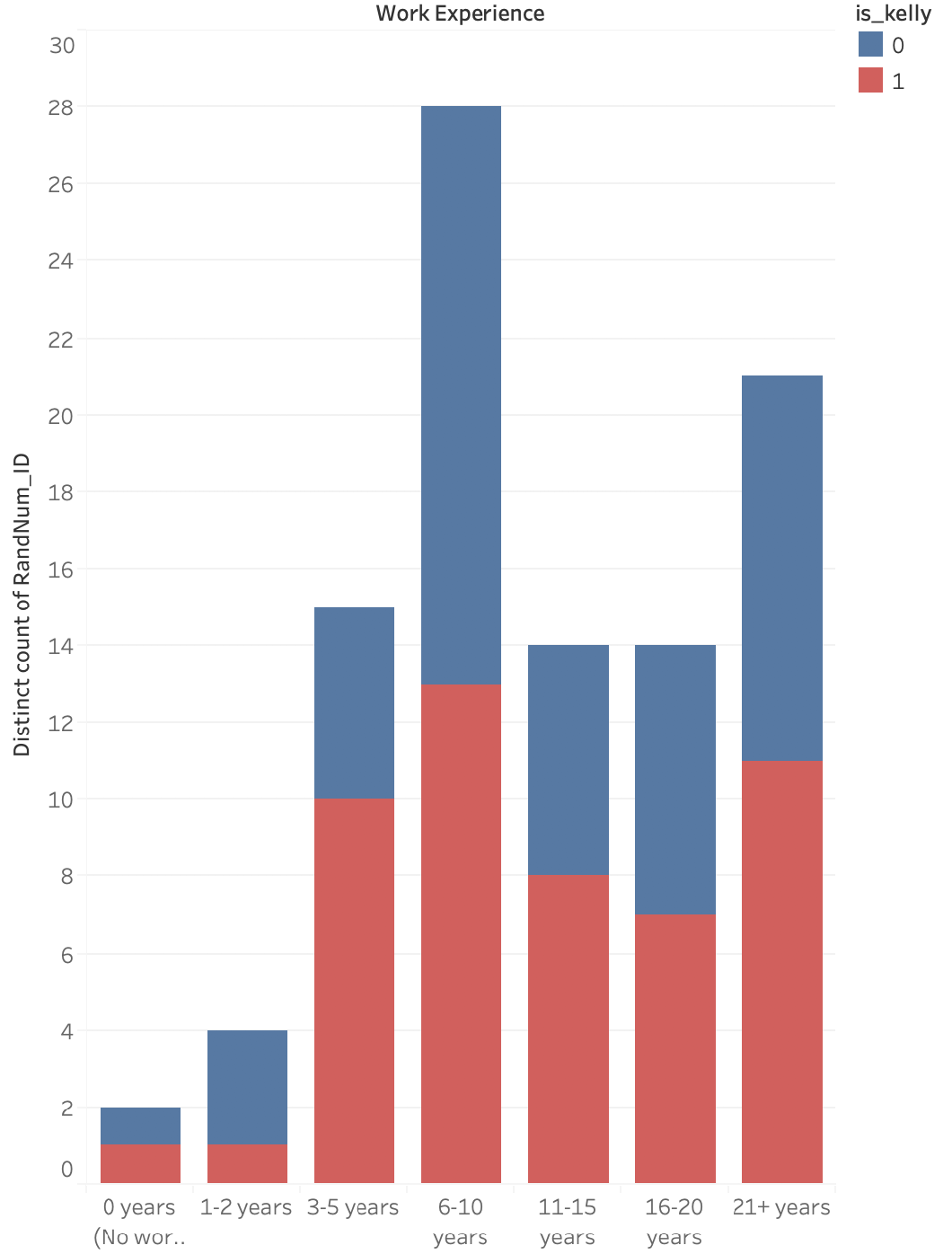
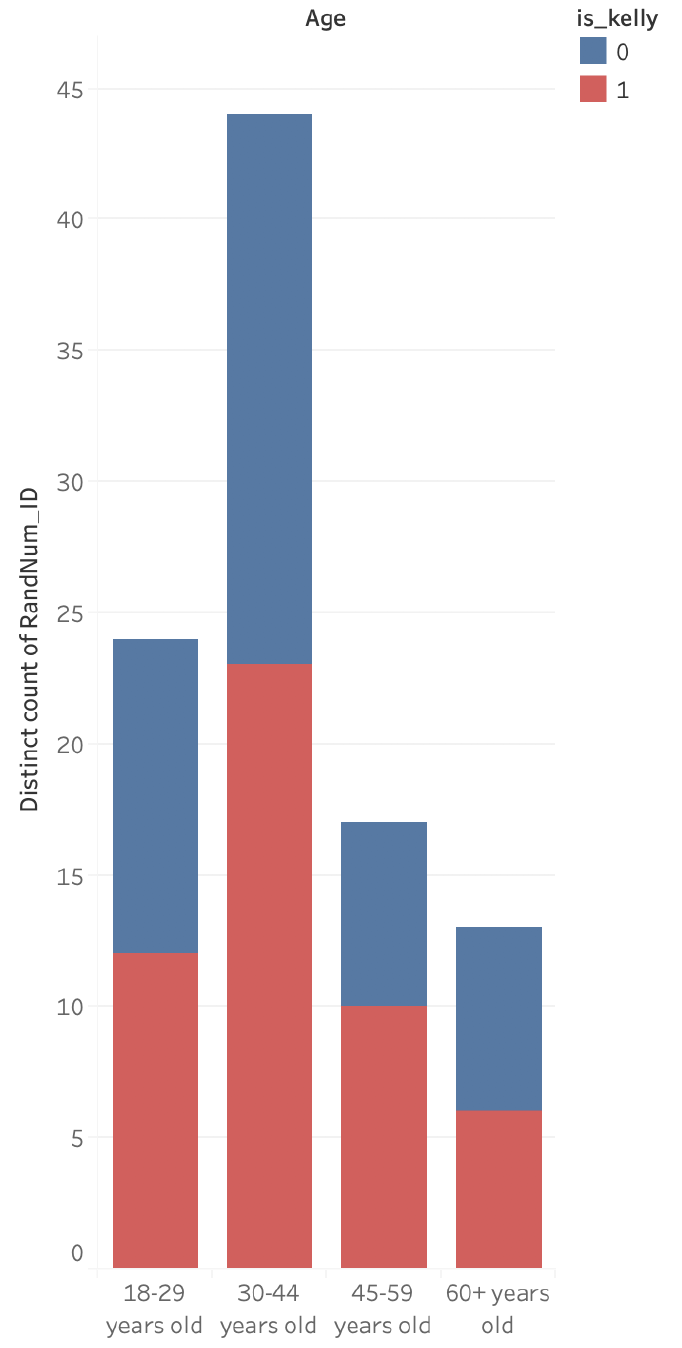
In order to identify any potential errors in the experiment design and get a preliminary understanding of the results, we conducted a pilot survey two weeks before rolling out our official experiment. For the pilot study, we were able to recruit 100 participants within a day through MTurk.

## Covariate Balance Check

Below are the covariate checks we ran based on the demographic inputs of the participants. Treatment group is represented in red, and control is represented in blue.

*Figure 2 - Race Distribution*  *Figure 3 - Education Distribution*

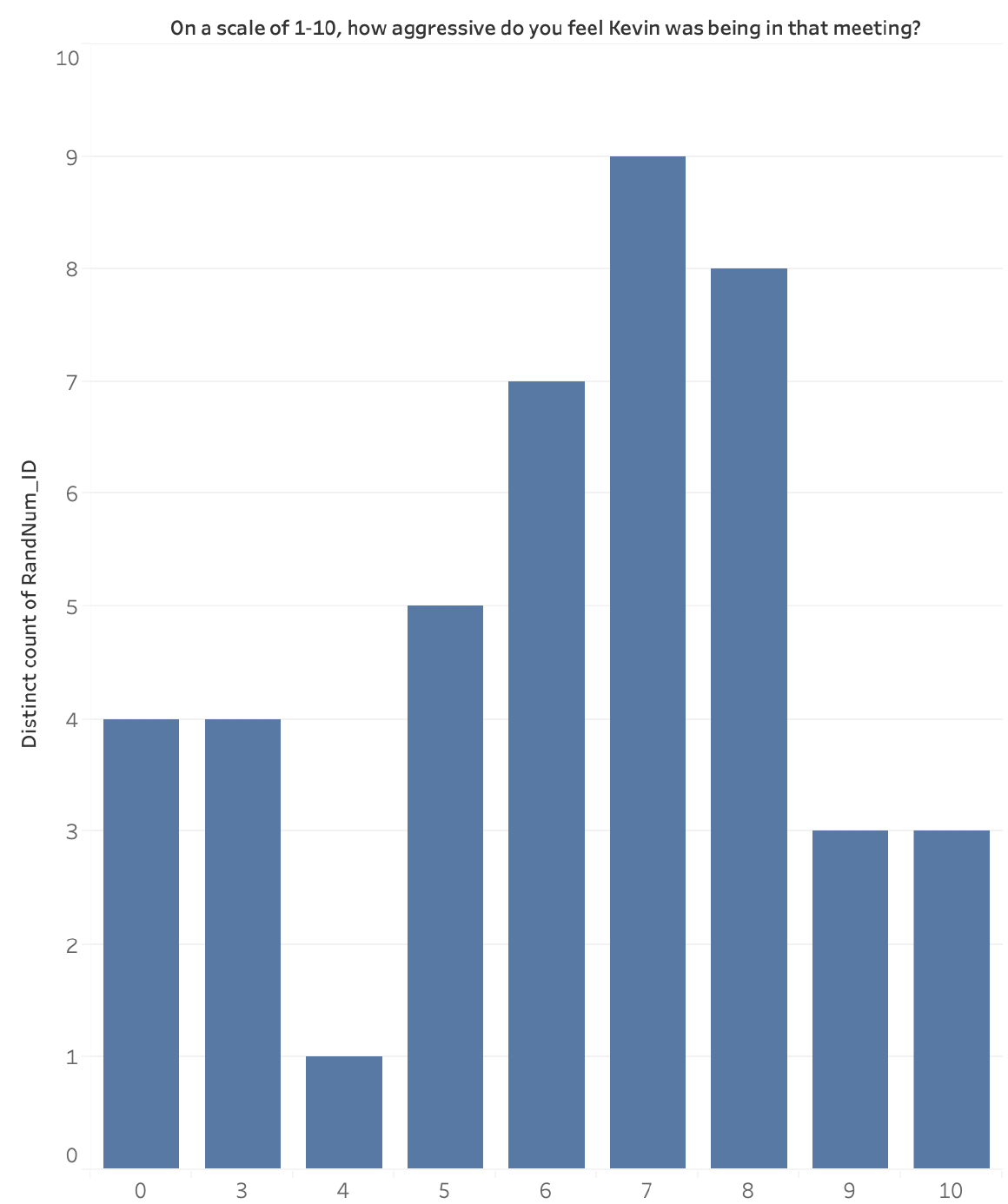
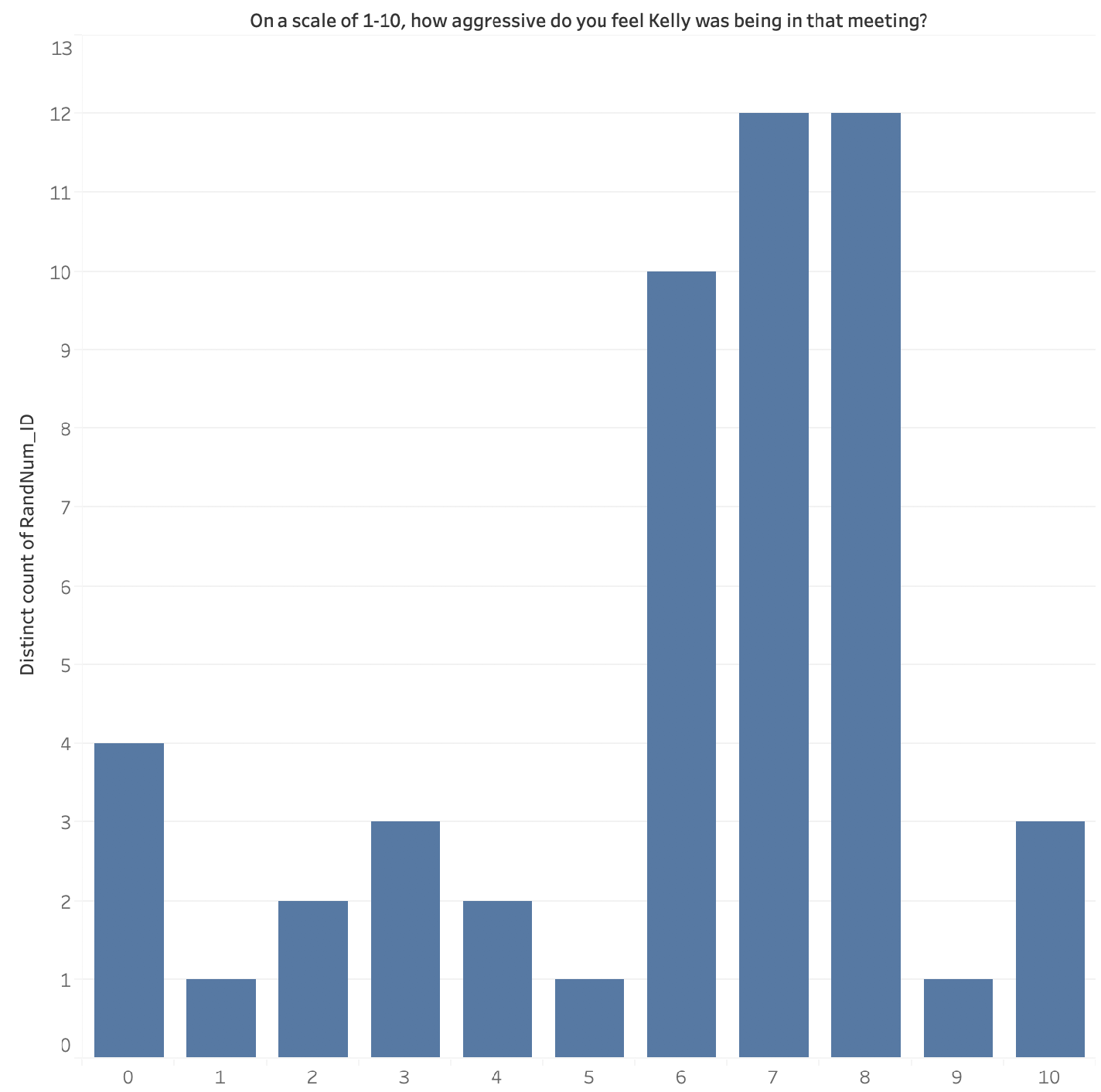
 

*Figure 4 - Work Experience Distribution*  *Figure 5 - Age Distribution*

We could tell by looking at the race distribution (Figure 2) that white population was the dominant. Although there’s a in-balance of the race group and may raise the question of generalizability of the study, this reflected the race distribution of North American society and thus was accepted in our survey. From Figure 4 we could see that the participants we recruited on MTurk generally have a few years of work experience. This raised a good sign for our study as our target audience was the working class.

## Statistical Power Check

We plotted the distribution of the aggressiveness scores of Kelly (treatment) and Kevin (control).

*Figure 6 - Aggressiveness Score Distribution (Control (Left) vs. Treatment (Right))*

Average treatment aggressiveness score was 6, with a standard deviation of 2.66, while the average control aggressiveness score was 6.06, with a standard deviation of 2.66 as well. From this, we were able to observe two things: One, that the mean scores actually ran directly counter to our hypothesis that the direction of the aggressiveness gender bias would run against women. Two, the strength of the effect in either direction was not statistically significant.

Although the directionality of the preliminary results was not favourable to our hypothesis, we nevertheless decided to run the results through a statistical power analysis in order to determine the sample size we would need to get in order to get a statistically significant effect. From a purely naive result not accounting for noncompliance, we determined that we would need a sample size slightly over 30,000. After eliminating non-compliers from consideration, that required sample size fell significantly, but remain high at 2,500+. In both cases, this exceeded the expected final sample size of 1,000, which meant that this preliminary check indicated we were overwhelmingly likely to return a null result. Even if we successfully found a statistically significant difference it would have pointed in a direction contrary to our initial hypothesis. Nevertheless, as part of the process we had laid out, we decided to follow through with our initial intentions in order to have a concrete final result to report upon for the benefit of our classmates.

## Noncompliance Check

For the pilot survey, our primary method of detecting noncompliance was by measuring the respondents’ survey completion time. A check of this survey completion time revealed that 25 respondents took less than 90 seconds to complete the survey. While we cannot know for sure that these respondents did not fully read the story, it indicates the possibility of noncompliance among respondents. Therefore, as a method to more accurately identify noncompliance, we added in the three story recall questions into the final survey.

# 

# 

# Final Study Results and Discussion

## 

## Sample Size and Noncompliance Check

For the final survey, we received a total of 832 completed responses. However, we used the responses for the story recall questions to get an understanding of noncompliance. For the first question, 91.7% of respondents answered correctly, 85.8% for the second question, and 86.1% for the third. In total, 647 (77.7%) respondents answered all three recall questions correctly.

We also used total response times as an indicator for noncompliance. Temporarily excluding those that took a long time to complete the survey (95% percentile), we calculated summary statistics and standard deviations for the response times, and ultimately deemed response times less than one standard deviation below the mean as noncompliance.

Using these two guidelines (recall question accuracy & response times), we removed non-compliers and resulted in a total of 613 responses (73.7% of total responses). For the majority of the following analysis, we will focus on and utilize the cleaned subset of the dataset with noncompliant respondents removed.

## 

## Respondent Demographics - Compilers Only

Below is information on the demographics of our respondents, using the cleaned dataset with compilers only.

|  |  |  |
| --- | --- | --- |
| **Age Group** | **#** | **%** |
| 18-29 | 159 | 25.9% |
| 30-44 | 247 | 40.3% |
| 45-59 | 137 | 22.3% |
| 60+ | 69 | 11.3% |
| Prefer not to say | 1 | 0.2% |

*Table 1 - Reported Age Group Count and Proportions*

|  |  |  |
| --- | --- | --- |
| **Gender** | **#** | **%** |
| Male | 300 | 48.9% |
| Female | 305 | 49.8% |
| Non-binary/third gender | 4 | 0.7% |
| Prefer not to say | 4 | 0.7% |

*Table 2 - Reported Gender Count and Proportions*

|  |  |  |
| --- | --- | --- |
| **Education** | **#** | **%** |
| Less than high school degree | 5 | 0.8% |
| High school graduate (high school diploma or equivalent including GED) | 131 | 21.4% |
| Associate degree in college (2-year) | 80 | 13.1% |
| Bachelor's degree in college (4-year) | 282 | 46.0% |
| Master’s degree | 94 | 15.3% |
| Doctoral degree | 13 | 2.1% |
| Professional degree (JD, MD) | 8 | 1.3% |

*Table 3 - Reported Educational Attainment Count and Proportions*

|  |  |  |
| --- | --- | --- |
| **Professional Work Experience** | **#** | **%** |
| 0 years (No work experience) | 14 | 2.3% |
| 1-2 years | 42 | 6.9% |
| 3-5 years | 94 | 15.3% |
| 6-10 years | 116 | 18.9% |
| 11-15 years | 94 | 15.3% |
| 16-20 years | 76 | 12.4% |
| 21+ years | 177 | 28.9% |

*Table 4 - Reported Working Experience Count and Proportions*

|  |  |  |
| --- | --- | --- |
| **Race** | **#** | **%** |
| White | 459 | 74.9% |
| Asian | 68 | 11.1% |
| Black or African American | 53 | 8.6% |
| American Indian or Alaska Native | 2 | 0.3% |
| Native Hawaiian or Pacific Islander | 1 | 0.2% |
| Multiple Races | 17 | 2.8% |
| Other | 13 | 2.1% |

*Table 5 - Reported Ethnicity Count and Proportions*

Overall, our respondent population leans white, relatively highly educated, and having a decent amount of work experience. While this does cause some concern about representativeness as this sample is likely not representative of the U.S. population, for our study we wanted to get a sample of adults with working experience (as our study focuses on gender bias in the workplace), hence why we decided to use MTurk over Berkeley XLAB.

Blocking - Sample Size

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age / Gender** | **Female** | **Male** | **Non-binary/third gender** | **Prefer not to say** |
| **18-29** | 109 | 109 | 1 | 1 |
| **30-44** | 164 | 198 | 3 | 1 |
| **45-59** | 90 | 75 | 0 | 1 |
| **60+** | 46 | 31 | 0 | 0 |
| **Prefer not to say** | 1 | 0 | 0 | 2 |

*Table 6 - Sample Size by Block - All Respondents*

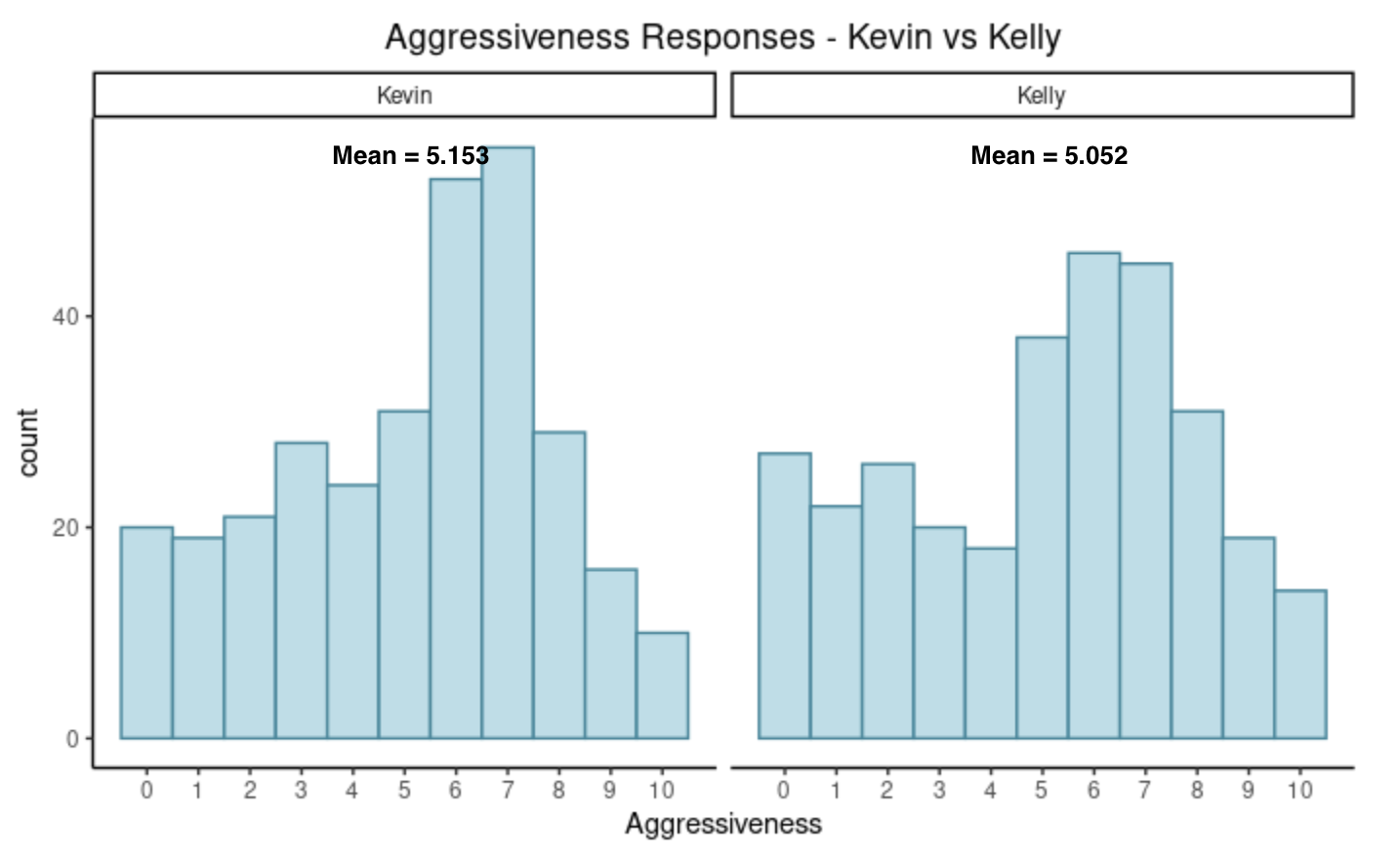
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age / Gender** | **Female** | **Male** | **Non-binary/third gender** | **Prefer not to say** |
| **18-29** | 82 | 75 | 1 | 1 |
| **30-44** | 112 | 131 | 3 | 1 |
| **45-59** | 69 | 67 | 0 | 1 |
| **60+** | 42 | 27 | 0 | 0 |
| **Prefer not to say** | 0 | 0 | 0 | 1 |

*Table 7 - Sample Size by Block - Compilers Only*

It is also important to look at sample sizes within each of our blocks. As we expected, we have very few responses for non-binary/third gender and for “Prefer not to say.” Within male and female, though, there is a relatively even split overall with female respondents leaning slightly older than male respondents. Looking at just male and female genders and excluding “Prefer not to say”, we were able to get more than 30 responses in each group. However, there are notably fewer responses for the 45-59 and 60+ age groups for both genders.

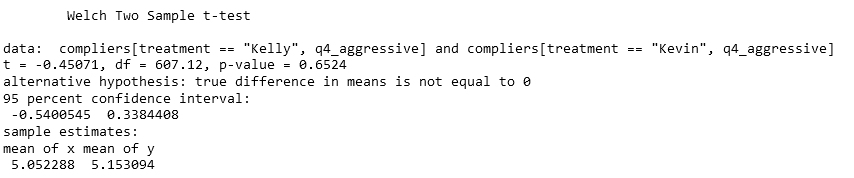
## Mean of Aggressiveness Responses - Treatment vs. Control

Before performing regression analysis, we examine the basic response distributions for our main outcome variable (aggressiveness of Kelly/Kevin). As seen in Figure 7, the mean of the responses is actually slightly lower for Kelly than for Kevin, contrary to our initial hypothesis. While it can be debatable whether the mean is the most accurate measure to compare, and different measures (e.g. median, % of responses rating aggressiveness 8 or higher, etc.) could potentially yield different results, we utilize the mean to get a basic understanding of our results prior to running regression analysis.



*Figure 7: Aggressiveness Response Distribution - Kevin vs Kelly, Compiler Subset*

To ascertain whether this difference between the treatment and control groups is statistically significant, we run an independent samples t-test (see Figure 8). In this test, we observed that the treatment group (protagonist is female) gave a mean aggressiveness rating of the protagonist of 5.05, while the control group (protagonist is male) gave a mean aggressiveness rating of 5.15. The mean values appear to be lower in the final group excluding noncompliers, which in turn is lower than the values before excluding noncompliers (treatment mean 5.534 vs control mean 5.535, respectively, which suggests an ITT of -0.0008 and a CACE of -0.0011). As with the pilot study, the point estimates ran contrary to our hypothesis, and the effect was not statistically significant (p-value 0.6524). This outcome was fully predicted by the results of our pilot study power analysis, which indicated that the study likely did not have sufficient power.

**

*Figure 8: Independent Samples T-Test Results on Complier Subset*

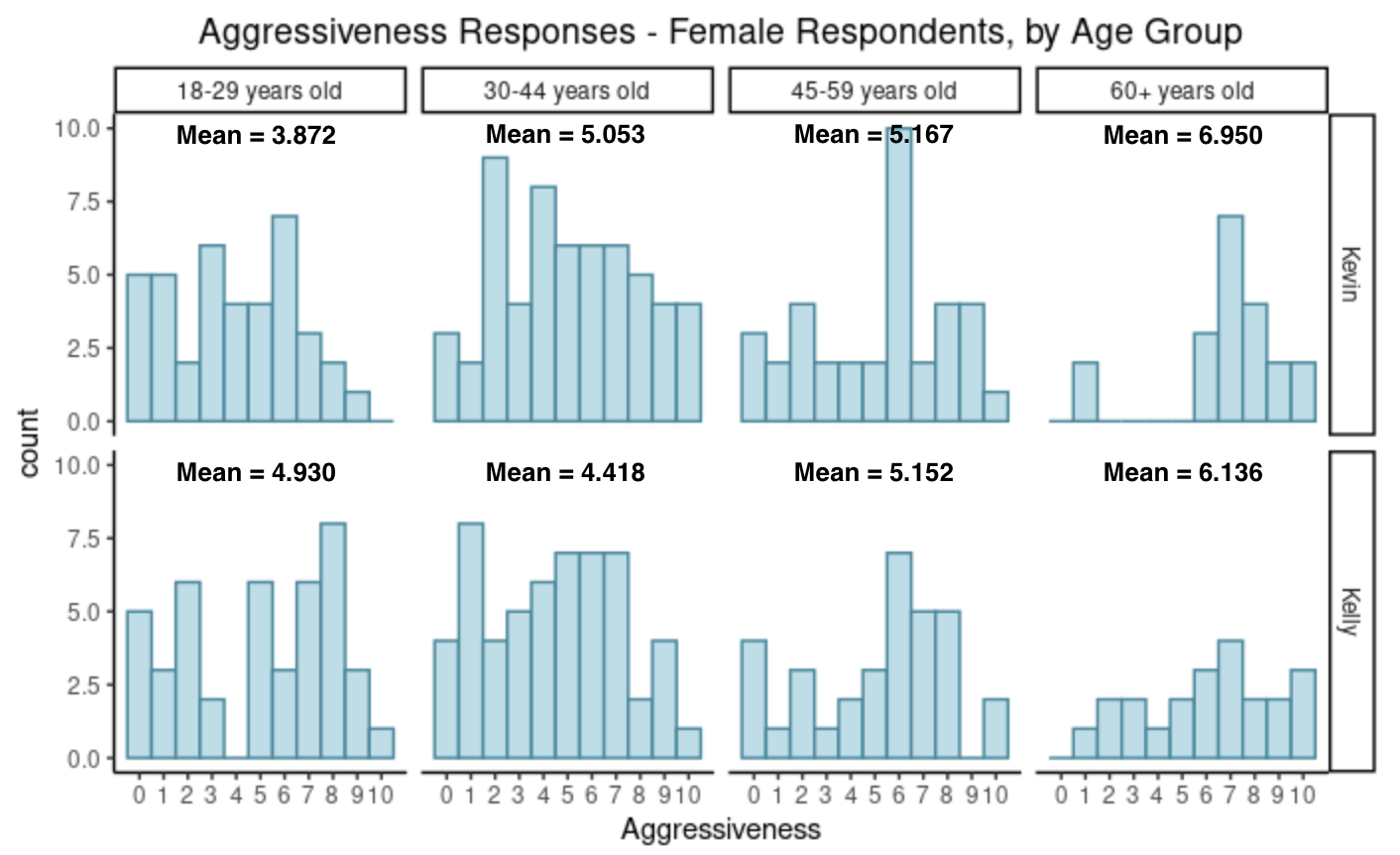
## 

## 

## High-Level Response Distributions - Compilers Only

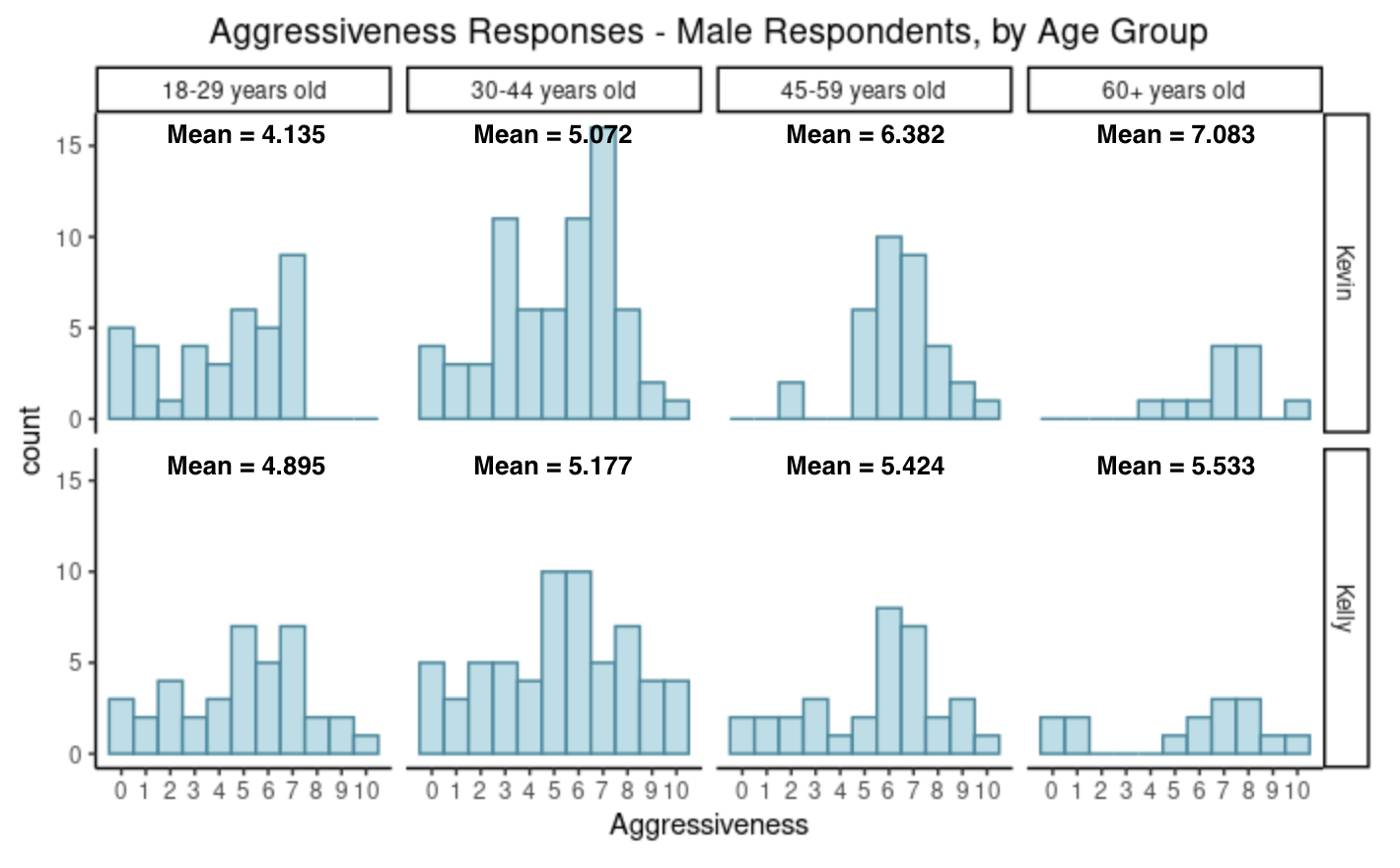
## 

## Continuing to take a basic look at response distributions prior to running regression analysis, below are response distributions and means for aggressiveness, broken out by our different blocks for our compilers dataset.



*Figure 9: Aggressiveness Responses by Block, Compiler Subset - Female Respondents*

As seen in Figure 9, the mean aggressiveness score female respondents in the 18-29 year old age group gave Kelly is higher than that for Kevin. For all other age groups, female respondents’ mean aggressiveness score for Kevin is higher than that for Kelly. However, looking just at the level of aggressiveness, female respondents in the 18-29 year old age group rated both Kevin and Kelly less aggressively (3-4 point range) overall than older age groups. In particular, the 60+ age group range rated both Kevin and Kelly higher, with mean scores above 6 for both.



*Figure 10: Aggressiveness Responses by Block, Compiler Subset - Male Respondents*

For male respondents, Figure 10 shows a similar pattern of respondents in older age groups overall having higher mean aggressiveness scores. Additionally, male respondents in the 18-29 and 30-44 age groups have higher mean aggressiveness scores for Kelly than for Kevin, while those in the 45-59 and 60+ age groups have higher mean aggressiveness scores for Kevin than for Kelly.

## Regression Analysis

Although the basic independent samples t-test of aggressiveness for Kevin vs Kelly did not indicate a statistically significant effect, we nevertheless move on to conduct regression analysis in order to see if we could narrow down the treatment effect a little more on the outcome variable. In this phase of the analysis, we build three models. The first is a baseline model, regressing the outcome against the treatment effect and the blocks as factors. The second model builds upon the first by including the aggressiveness rating of the manager (Sean) as a means to control for variations between respondents (i.e. some people may have a propensity to give a base score). Finally, the third model builds upon the second by filtering for non-compliers beforehand, as well as incorporating the measurements of whether the respondent felt the actions were due to factors internal or external to the characters. In all cases, we use heteroscedasticity-robust standard errors (type HC0). The full output comparing these three models can be found in Appendix C.

P-values are also reported upon for these, but the sheer number of comparisons raises the spectre of a fishing expedition. While more rigorous approaches like the Bonferroni Correction can let us account for this, we are primarily interested in gauging the relative strength of the block effects, and not looking to draw any hard conclusions. Hence, for the purposes of the analysis below, the multiple comparisons problem will be addressed simply by only considering effects at the 99% confidence interval to be of statistical significance.

Concerning the choice of variable inputs, there was some debate amongst our group as to whether the inclusion of the manager (Sean) aggression variable to capture variance between subjects constituted a bad control. On the one hand, the treatment should have been independent of the manager, as it was restricted to the gender of the protagonist. On the other hand, it constitutes something that is measured after the treatment. We decided to include it in the final models, albeit with the awareness that the appropriateness of this variable is debatable.

The first model has a treatment variable point estimate of nearly zero (0.004), with a standard error of 0.188 leading to a p-value of 0.984. In this first iteration, the effect was virtually undetectable. Still, the coefficients of the different blocks indicate some evidence that certain groups did give higher aggressiveness scores irrespective of whether they were in treatment, with the Female 60+, Male 30-44, Male 45-59, and Male 60+ group exhibiting higher scores at a 99% confidence interval. There were other blocks with standard errors rising to the 99% CI, but the extremely small sample size of those blocks means the veracity of their reported effect sizes is dubious at best.

After this, the second model shows a sizable improvement in the point estimate of the treatment variable, though the standard error means that the effect still fails to rise to statistical significance (p = 0.385). As with the t-test, the direction of the estimate indicates that the score was seen to be higher for control rather than treatment. This improvement in precision was exclusively the result of the incorporation of the aggressiveness score of the manager, which appears to have served its intended purpose of accounting for the respondents’ baseline values, with a point estimate of 0.434 and a standard error of 0.039. As for scores by blocks, the same groups exhibited higher scores of statistical significance as the first model (Female 60+, Male 30-44, Male 45-59, and Male 60+), though the female 60+ group’s point estimate was a fair bit lower (1.241 versus 1.607 in the first model).

Finally, the exclusion of identified non-compliers in the third model yielded some mixed results. The treatment effect point estimate increased further, to 0.204. However, the standard error also went up to 0.208, meaning that although the treatment effect moved closer to statistical significance overall (p = 0.325), the regression modeling approach ultimately also failed to provide sufficient evidence for us to reject the null hypothesis. The inclusion of the questions about action attribution also did not yield any statistically distinguishable results. Interestingly, the strength of the block effects increases for the Female 60+, Male 45-59, and Male 60+ groups, though the reduced sample size of the latter renders the increase in group effect size questionable. As well, the Male 30-44 block effect has diminished to the point that it is no longer statistically significant at the 99% confidence level.

Limitations

There are several hypotheses for this lack of observed statistical significance, which we believe are related to the limitations inherent in the means by which this study was delivered. While the point estimates indicate that the anti-female gender bias simply will not manifest given more data, these conjectural explanations will assume that the point estimates will point in the expected direction given a revised experiment.

## 

## 

## Generalizability and Platform Choice

Our respondent group was drawn exclusively from Mechanical Turk, and the timeframes in which we drew respondents were fairly short. This raises questions about the generalizability of the study, as well as restricting the respondents to a small demographic group of relatively tech-savvy, predominantly white subset. The nature of the study conducted means these results are primarily Anglocentric in nature, and may not reflect views held in other cultures as well. The choice to use MTurk was primarily motivated by the difficulties associated with collecting in-person field data in the midst of a pandemic while scattered across three different continents, but it does raise the possibility that more rigorous data collection methods may have yielded a result more consistent with the literature review that led to this experiment.

Noncompliance Measurement

Another consequence of the choice of platform was the fact that we had imperfect ability to measure noncompliance. While we did collect metrics both visible and invisible to the end users in order to measure and control for noncompliance, we simply have no guarantee that the respondents surveyed actually received the treatment in the intended method. This appears to be an inherent limitation to recruiting respondents in an online platform - given more resources, we could take more effort to ensure compliance on the part of the respondents by taking actions like asking followup questions to test for understanding.

Delivery Mechanism

In the context of this experiment, we delivered the treatment via text. Compared to other communication mediums like audio and video, this approach is relatively abstract and sterile, lacking in the sorts of cues that we would associate with gender such as vocal inflection, appearance, and clothing. This hypothesis is related to the question of the treatment dosage, as employing a richer medium may have made the gender differences sufficiently prominent as to trigger the effect. Although such an approach was out of scope of this project given time and resource availability considerations, this represents another avenue of exploration in future research on gender bias.

Treatment Dosage

In this experiment, the only thing that varied was the gender of the protagonist, along with the pronouns used to identify them. Writing is an inherently qualitative exercise, which makes it difficult to determine the dosage applied. It is possible, therefore, that the dosage was simply too low for there to be an effect. For instance, perhaps the use of additional punctuation could have been used to magnify the aggression to the point that differences could manifest.

Other substitutes for the dosage metric may include, but is not limited to the number of references to the character’s gender, or references to gender-specific attributes, or other attributes. In these instances, perhaps highlighting masculinity or femininity could have brought out a stronger difference between treatment and control. In any case, the difficulty in measuring dosage has made it difficult to know what sort of dosage would be required to achieve a measurable effect. Another aspect of this dosage hypothesis is the length for which the respondents know the characters, as the story essentially exists in a vacuum devoid of any context. If the respondents knew the characters for a longer period of time over more interactions, a difference in perception arising from gender might manifest.

## 

## 

## Lessons for Future Iterations of Experiment

Based on the literature, we do not believe that our failure to reject the null hypothesis stemmed from the nonexistence of the effect - rather, our measurement strategy and treatment administration strategy do appear, in hindsight, to have been insufficient to bring out the effect. If we were given unlimited time and resources to bring out the gender aggressiveness bias in the future, we would not only conduct this experiment with a much larger sample size, but with a better understanding of the dosage of the treatment. This hypothetical future experiment would also make use of platforms where compliance can be assured. We would also carefully target the medium requirements to allow the bias to manifest, and target those requirements more precisely.

# Conclusion

While it is somewhat disappointing that our experiment did not yield significant results and that we did not get results that reproduce the effects seen in existing literature surrounding gender bias in the workplace, it is also a relief that a simple 500 (or so) word story did not generate statistically significant evidence of gender bias - if bias were that trivial to bring to the forefront, it would raise serious questions either about the methodology of our study, or the state of gender equality in the world. This process proved to be an invaluable learning process for us, as we were able to research, devise, and implement an experiment, building and utilizing an entire pipeline along the way. This learning was a small consolation for the inability of the experiment to detect a statistically significant effect. This experiment also gave us an important starting point for future iterations, in which we intend to find ways to better quantify and increase the treatment dosage and seek a pool of respondents over whom we can better ensure compliance in order to create an experiment with the highest possible statistical power from the beginning.

# 

# References

[1] Women in the Workplace 2020 <https://www.mckinsey.com/featured-insights/diversity-and-inclusion/women-in-the-workplace>

[2] How Gender Bias Corrupts Performance Reviews, and What to Do About It

<https://hbr.org/2017/04/how-gender-bias-corrupts-performance-reviews-and-what-to-do-about-it>

[3] When Women Are Called 'Aggressive' At Work

<https://www.forbes.com/sites/nextavenue/2018/08/28/when-women-are-called-aggressive-at-work/?sh=44dac7957bc8>

[4] Top Names Over the Last 100 Years

<https://www.ssa.gov/oact/babynames/decades/century.html>

[5] GloVe: Global Vectors for Word Representation

<https://nlp.stanford.edu/projects/glove/>

# Appendix A - Survey and Story

**CONTROL**

**First, we would like to ask a few demographic questions. Please answer the following questions before proceeding with the survey:**

1. What is your gender?
   1. Male
   2. Female
   3. Non-binary/third gender
   4. Perfer not to say
2. What is your age?
   1. 18-29 years old
   2. 30-44 years old
   3. 45-59 years old
   4. 60+ years old
   5. Prefer not to say
3. What is the highest level of school you have completed or the highest degree you have received?
   1. Less than high school degree
   2. High school graduate (high school diploma or equivalent including GED)
   3. Associate degree in college (2-year)
   4. Bachelor's degree in college (4-year)
   5. Master's degree
   6. Doctoral degree
   7. Professional degree (JD, MD)
4. How many years of professional work experience do you have?
   1. 0 years (No work experience)
   2. 1-2 years
   3. 3-5 years
   4. 6-10 years
   5. 11-15 years
   6. 16-20 years
   7. 21+ years
5. Choose one or more races that you consider yourself to be:
   1. White
   2. Black or African American
   3. American Indian or Alaska Native
   4. Asian
   5. Native Hawaiian or Pacific Islander
   6. Other

**Please read the following text. Afterwards, you will be asked a couple of questions based on the content of the text.**

Kevin is a senior employee with Geospatial Analytics, a consulting firm. In the past year, he has been responsible for completing several major projects, which have collectively brought in a sizable amount of money. After spending long hours working on the projects, Kevin did some research on the compensation he was receiving at the firm. He found that someone with his skillset working a similar number of hours should be receiving about 15% higher pay. With this information in mind, Kevin booked a brief meeting with his boss, Sean, with the intention of asking for a raise of that amount.

At the meeting, Kevin started with a recap of the work done over the past year. “As you know, I’ve contributed to all the major projects that we’ve turned around this year. I’ve spent a lot of time meeting with the respective team leads to make sure that we’ve scoped our work properly, and because of my accurate forecasting, the projects I’ve been assigned to were overall on time and ten percent below budget. I’ve even managed to deliver extra value by routinely identifying future avenues of automation to help cut down on future maintenance work for the clients.”

This point was well received by Sean. “Yes, all of that is true”, he said. “The amount of effort and attention to detail that you’ve put into your projects has been noticed by senior management, and we are appreciative of everything that you’ve done for Geospatial Analytics in the past year.”

With that positive response, Kevin got to the point. “Thank you. In the coming year, we will be taking on a lot more projects than ever before, and we will need to expand our headcount to ensure we maintain the same quality as what we delivered with our past projects. As we do this, we’re going to need talented team leads to bring the new colleagues up to speed. I’m confident I can step up and help with that process.”

“Given all that I’ve accomplished over the past year, and the fact that the scope of my responsibilities are set to expand, I’d like to ask for a 15% increase in my pay.”

Sean defaulted to giving a noncommittal answer. “I appreciate all the hard work and long hours you’ve put in. Unfortunately, we just don’t have the budget for a 15% raise for you this year. Do you think you can wait another year before resubmitting this request?

Kevin decided to press the issue, despite Sean’s less than enthusiastic response. "Then I suggest you re-examine your budget. I’ve done my research, and given my responsibilities, my pay should be 15% higher. If I’m going to deliver the best results for you, I need to feel like I’m respected by the organization, and based on my research, I think being very reasonable here.”

Sean cut the discussion short, responding “as I said, we don’t have the budget, and as much as I like you, this just isn’t going to happen. I’m sorry, but I can’t give you what you’re asking for. Now, if that’s all, then we have the quarterly meeting to prepare for.”

You are a mutual friend of Kevin and Sean, and have heard the story from both sides. In order to determine whose side you want to take in this disagreement, you want to assess how aggressive the two were being, and whether their actions were because of factors external to them (e.g. having a bad day), or internal (e.g. character traits):

1. What is the salary increase that Kevin asked for?
   1. 5%
   2. 10%
   3. 15%
   4. Was not mentioned in the text
2. How did Kevin start the conversation?
   1. Recapped the work he has done over the past year
   2. Got straight to her request
   3. Recognized the entire team for their support
3. What was Sean's reasoning for denying Kevin’s request?
   1. Kevin’s performance doesn't warrant a raise
   2. Does not have the budget for salary increases
   3. Strict company policy on timing of salary increases
4. On a scale of 0-10, how aggressive do you feel Kevin was being in that meeting?
5. On a scale of 0-10, how aggressive do you feel Sean was being in that meeting?
6. Did you feel that Kevin’s actions were driven more by internal factors or by external factors?
   1. Internal factors
   2. External factors
7. Did you feel that Sean’s actions were driven more by internal factors or by external factors?
   1. Internal factors
   2. External factors

**TREATMENT**

**First, we would like to ask a few demographic questions. Please answer the following questions before proceeding with the survey:**

1. What is your gender?
   1. Male
   2. Female
   3. Non-binary/third gender
   4. Perfer not to say
2. What is your age?
   1. 18-29 years old
   2. 30-44 years old
   3. 45-59 years old
   4. 60+ years old
   5. Prefer not to say
3. What is the highest level of school you have completed or the highest degree you have received?
   1. Less than high school degree
   2. High school graduate (high school diploma or equivalent including GED)
   3. Associate degree in college (2-year)
   4. Bachelor's degree in college (4-year)
   5. Master's degree
   6. Doctoral degree
   7. Professional degree (JD, MD)
4. How many years of professional work experience do you have?
   1. 0 years (No work experience)
   2. 1-2 years
   3. 3-5 years
   4. 6-10 years
   5. 11-15 years
   6. 16-20 years
   7. 21+ years
5. Choose one or more races that you consider yourself to be:
   1. White
   2. Black or African American
   3. American Indian or Alaska Native
   4. Asian
   5. Native Hawaiian or Pacific Islander
   6. Other

**Please read the following text. Afterwards, you will be asked a couple of questions based on the content of the text.**

Kelly is a senior employee with Geospatial Analytics, a consulting firm. In the past year, she has been responsible for completing several major projects, which have collectively brought in a sizable amount of money. After spending long hours working on the projects, Kelly did some research on the compensation she was receiving at the firm. She found that someone with her skillset working a similar number of hours should be receiving about 15% higher pay. With this information in mind, Kelly booked a brief meeting with her boss, Sean, with the intention of asking for a raise of that amount.

At the meeting, Kelly started with a recap of the work done over the past year. “As you know, I’ve contributed to all the major projects that we’ve turned around this year. I’ve spent a lot of time meeting with the respective team leads to make sure that we’ve scoped our work properly, and because of my accurate forecasting, the projects I’ve been assigned to were overall on time and ten percent below budget. I’ve even managed to deliver extra value by routinely identifying future avenues of automation to help cut down on future maintenance work for the clients.”

This point was well received by Sean. “Yes, all of that is true”, he said. “The amount of effort and attention to detail that you’ve put into your projects has been noticed by the senior management, and we are appreciative of everything that you’ve done for Geospatial Analytics in the past year.”

With that positive response, Kelly got to the point. “Thank you. In the coming year, we will be taking on a lot more projects than ever before, and we will need to expand our headcount to ensure we maintain the same quality as what we delivered with our past projects. As we do this, we’re going to need talented team leads to bring the new colleagues up to speed. I’m confident I can step up and help with that process.”

“Given all that I’ve accomplished over the past year, and the fact that the scope of my responsibilities are set to expand, I’d like to ask for a 15% increase in my pay.”

Sean defaulted to giving a noncommittal answer. “I appreciate all the hard work and long hours you’ve put in. Unfortunately, we just don’t have the budget for a 15% raise for you this year. Do you think you can wait another year before resubmitting this request?

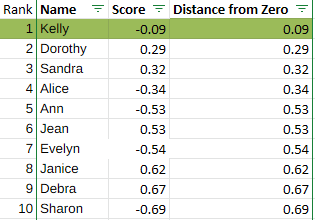
Kelly decided to press the issue, despite Sean’s less than enthusiastic response. "Then I suggest you re-examine your budget. I’ve done my research, and given my responsibilities, my pay should be 15% higher. If I’m going to deliver the best results for you, I need to feel like I’m respected by the organization, and based on my research, I think being very reasonable here.”

Sean cut the discussion short, responding “as I said, we don’t have the budget, and as much as I like you, this just isn’t going to happen. I’m sorry, but I can’t give you what you’re asking for. Now, if that’s all, then we have the quarterly meeting to prepare for.”

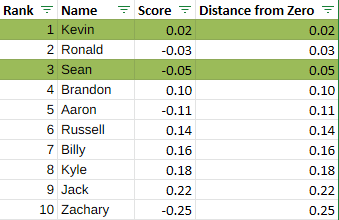
You are a mutual friend of Kelly and Sean, and have heard the story from both sides. In order to determine whose side you want to take in this disagreement, you want to assess how aggressive the two were being, and whether their actions were because of factors external to them (e.g. having a bad day), or internal (e.g. character traits):

1. What is the salary increase that Kelly asked for?
   1. 5%
   2. 10%
   3. 15%
   4. Was not mentioned in the text
2. How did Kelly start the conversation?
   1. Recapped the work she has done over the past year
   2. Got straight to her request
   3. Recognized the entire team for their support
3. What was Sean's reasoning for denying Kelly's request?
   1. Kelly's performance doesn't warrant a raise
   2. Does not have the budget for salary increases
   3. Strict company policy on timing of salary increases
4. On a scale of 0-10, how aggressive do you feel Kelly was being in that meeting?
5. On a scale of 0-10, how aggressive do you feel Sean was being in that meeting?
6. Did you feel that Kelly’s actions were driven more by internal factors or by external factors?
   1. Internal factors
   2. External factors
7. Did you feel that Sean’s actions were driven more by internal factors or by external factors?
   1. Internal factors
   2. External factors

# Appendix B - Top Sentiment Neutral Names by Gender

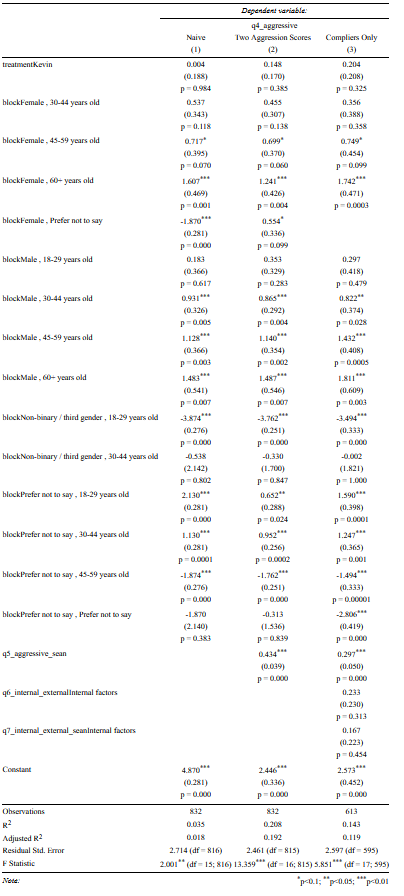


*Table B-1: Most Sentiment-Neutral Treatment (Female) Names*



*Table B-2: Most Sentiment-Neutral Control (Male) Names*

# Appendix C - Regression Models Full Output



*Table C: Regression Model Coefficients and Summary*