Analyzing Empathy Bias Towards One's Gender

David Jiang

Introduction

"Am I the Jerk"[1] is a popular internet forum where individuals describe an experience they had, and people discuss whether the individual was the jerk in that situation and to what degree. This report analyzes the responses to such scenarios to investigate whether or not respondents will rate individuals differently if they identify with the individual's gender. No particular such relationship was found in analysis.

Background

Data consists of the responses to variations of a survey, totalling 641 responses across all surveys. Each response includes the time of the response, respondent's academic level, the political affiliation of the adults the respondent grew up with, the respondent's political affiliation, some measure of their religiousness, respondent gender, and then 14 scenarios in which the respondent was asked to judge to what degree the person in the scenario was a jerk, ranging from "Not a jerk", "Slightly a jerk", and "Strongly a jerk."

The primary variation between surveys was in changing the genders of all people in each scenario. For example, while one survey may include a question on a man and his son, another survey may have changed it to a woman and her daughter. Surveys also varied in how they measured the respondent's religiousness. Some asked how religious the respondent was, some instead asked respondents to rate how spiritual they were, and one instead asked how often the respondent attended church. One survey included a priming question asking respondents whether they considered themselves compassionate or not. Finally, one survey (the one that asked how often respondents attended church) also asked how often their parents attended church, how often they watched sports, and how often their parents watched sports.

Due to technical issues with how surveys were distributed, some respondents left survey questions unanswered. However, missing responses made up at most 9% of the data for any given variable this analysis was concerned with. Due to the relatively low amount of missing data and the adverse effects imputation may have had on PCA analysis and hypothesis testing, it was decided to drop missing observations.

Another issue presented itself in certain questions not having the gender of those involved swapped correctly and instead having conflicting pronouns describing those in the scenario. This would make it impossible to determine what gender the respondent thought the people in the scenario were, and thus, whether or not they may have responded differently due to identifying with those in the scenario. These questions were thus not included in hypothesis testing.

[1] Terminology sanitized for this report.

All variables were transformed into numeric types for use in analyses. Scenario responses had their answers mapped to 0, 1, and 2 for “Not a Jerk,” “Slightly a Jerk,” and “Strongly a Jerk,” respectively. This scale was not 0 centered, as “Slightly a Jerk” may not necessarily be interpreted to be a neutral position. Gender responses were one-hot encoded according to each option provided, generating up to 4 dummy variables if non-binary/other or “prefer not to say” responses were present.

Some variables provided trouble in their “other” encodings. For academic cohort, while Freshman through Senior could be encoded from 0 to 3 (roughly how many years they had spent in school), Graduate Students would encompass anywhere from 4 to 9, depending on if they worked towards their Master’s degree part time. Meanwhile, some surveys only provided an “Other” option, making it entirely unclear where the respondent may fall within that distribution. Graduate Students were encoded as 4 to keep the data ordinal at the very least, while “Other” was encoded as -1. Due to the very low number of responses that fell into these categories, these encodings should have little impact on analyses.

For age, one option provided was “50+”. While other responses were exact years, “50+” may encompass a variety of possible ages. These few responses were simply encoded as 50 – as before, the lack of such responses means that such an encoding should not have adverse effect on analysis.

For political affiliation (as well as the one question on the political affiliation of the adults a respondent grew up with), responses ranged from “Strongly Liberal” to “Strongly Conservative,” with an option for “Don’t know / It’s complicated.” While typical responses could be put on a scale from -2 to 2, “Don’t know” could not be. It was decided instead to treat these responses missing. These made up a small minority of the data and imputation with the mean may have led to artificial clustering of data when performing PCA, so instead they were ignored.

Finally, questions that asked the frequency of events (such as how often a respondent attended church or watched sports) only allowed for the subjective responses of “Never,” “Frequently,” and “Often.” Since the difference between “Frequently” and “Often” was unknown, these were both coded as 1s while “Never” was encoded as 0.

Findings

Principal Component Analysis:

To investigate covariance of responses to each question, PCA was performed on the dataset of each survey. Results are summarized in the figures below. Note that the vectors of most original variables are omitted for clarity. Only the vectors corresponding to each gender option and the largest vectors that point in unique directions are retained. Original variables are named for the demographic metrics surveys measured, or denoting an “Am I the Jerk” question following questions about the respondent. Such variables follow the naming convention “q[question number][gender(s) of the individual(s) affected in the scenario, if specified]”.

Dummy variables for each option provided for the gender question were named with the convention “gender_[option]”.

Scree plots and PCA scatterplots where the data are collapsed to the two most significant principal components begin on the next page.

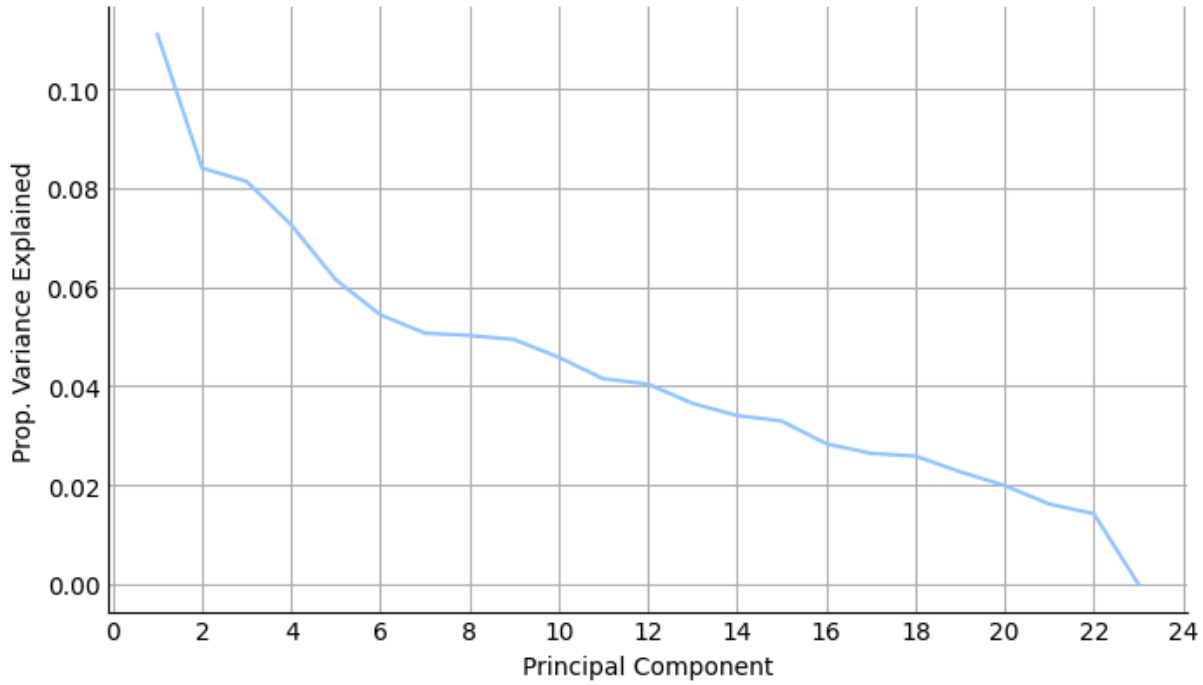

Figure 1. Scree Plot of PCA for Dataset 2024

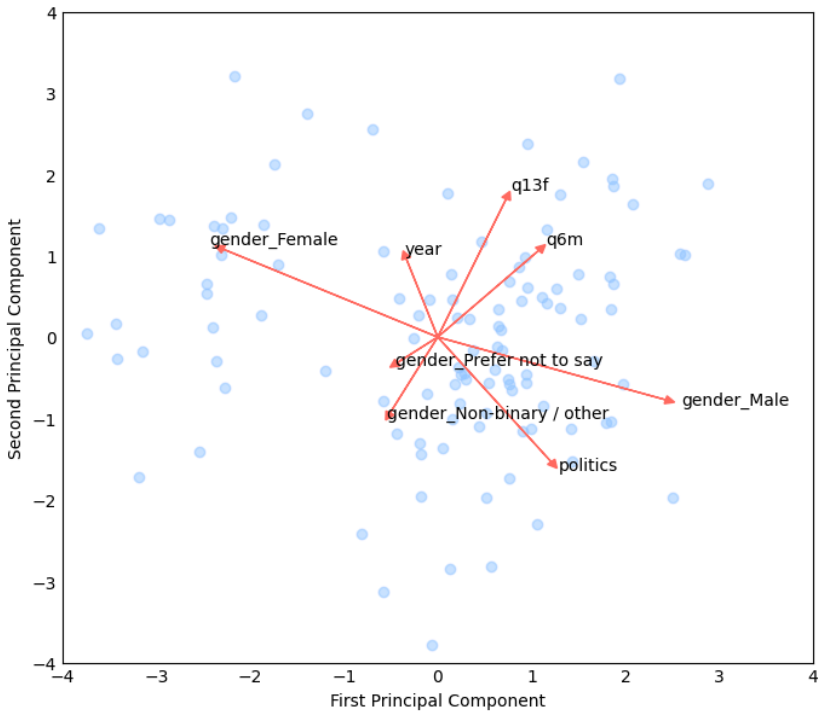

Figure 2. Dataset 2024 Reduced to First Two Principal Components

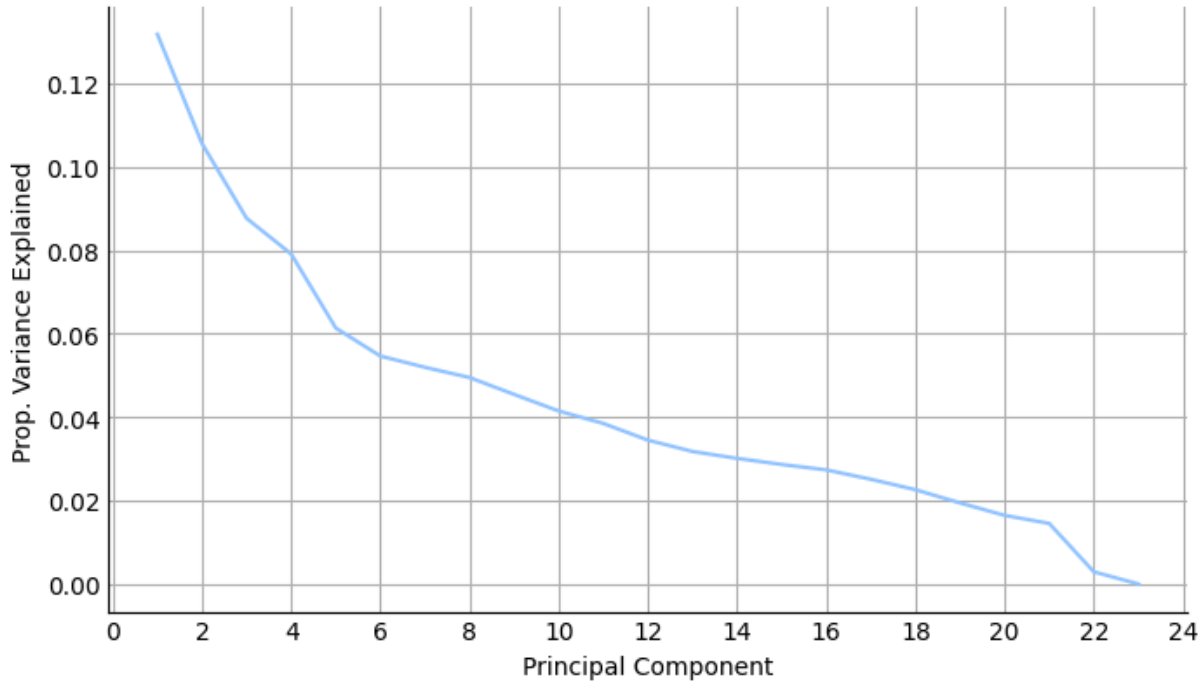

Figure 3. Scree Plot of PCA for Dataset "Fardina"

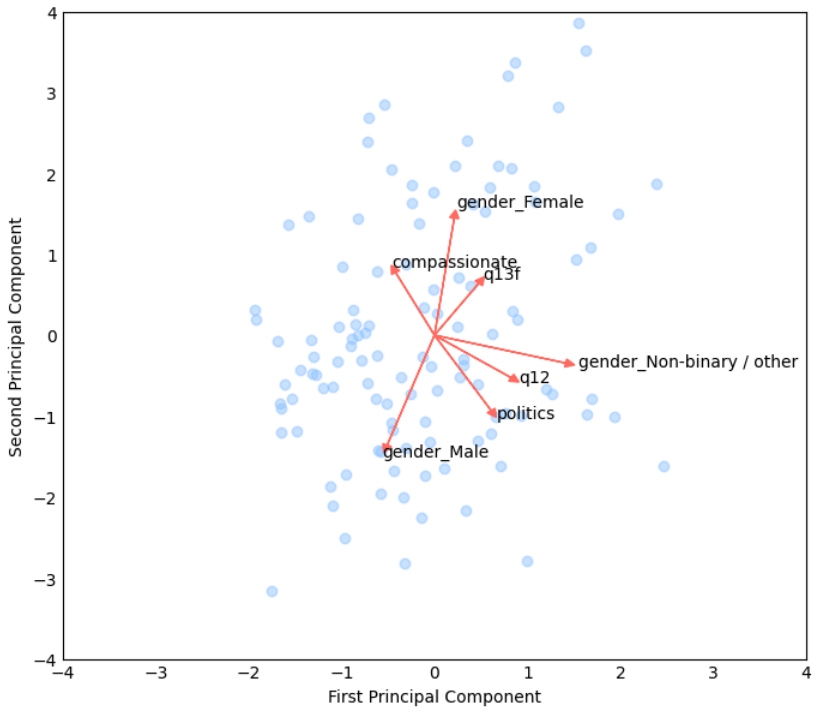

Figure 4. Dataset "Fardina" Reduced to First Two Principal Components

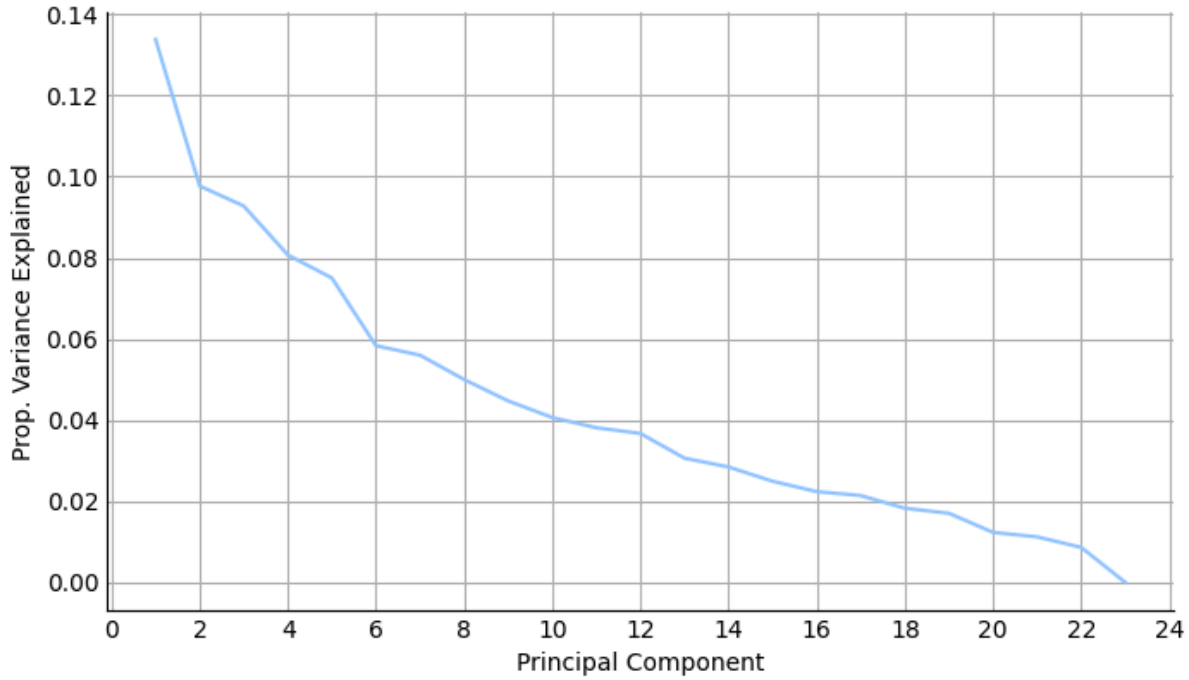

Figure 5. Scree Plot of PCA for Dataset "Max"

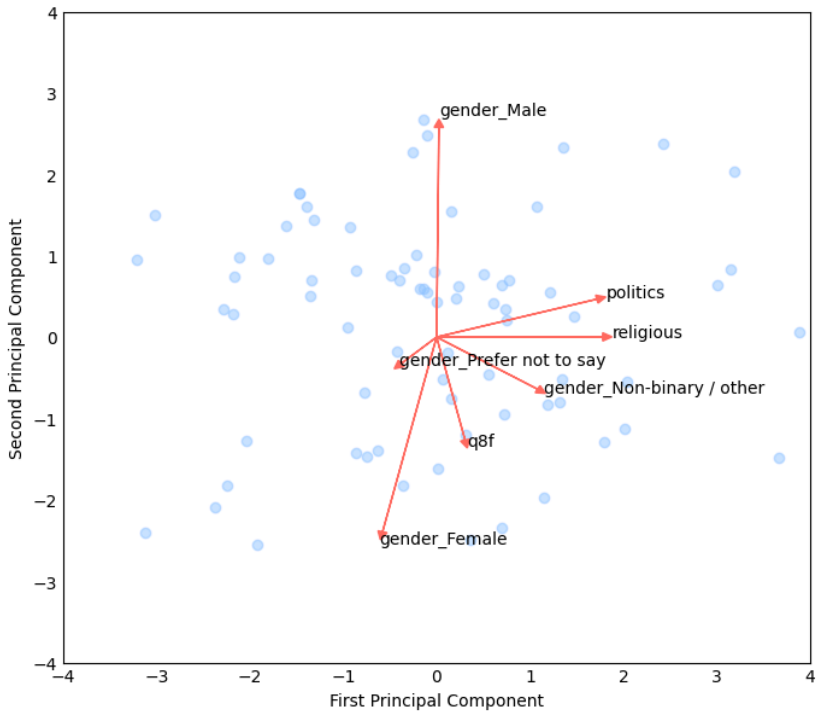

Figure 6. Dataset "Max" Reduced to First Two Principal Components

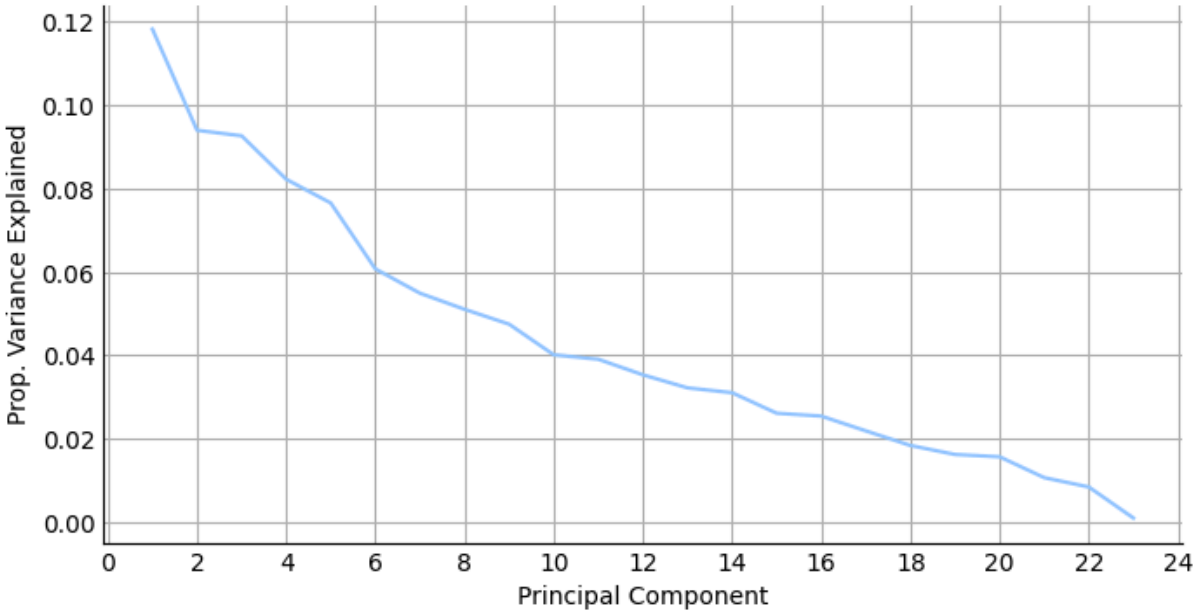

Figure 7. Scree Plot of PCA for Dataset Spring 2025

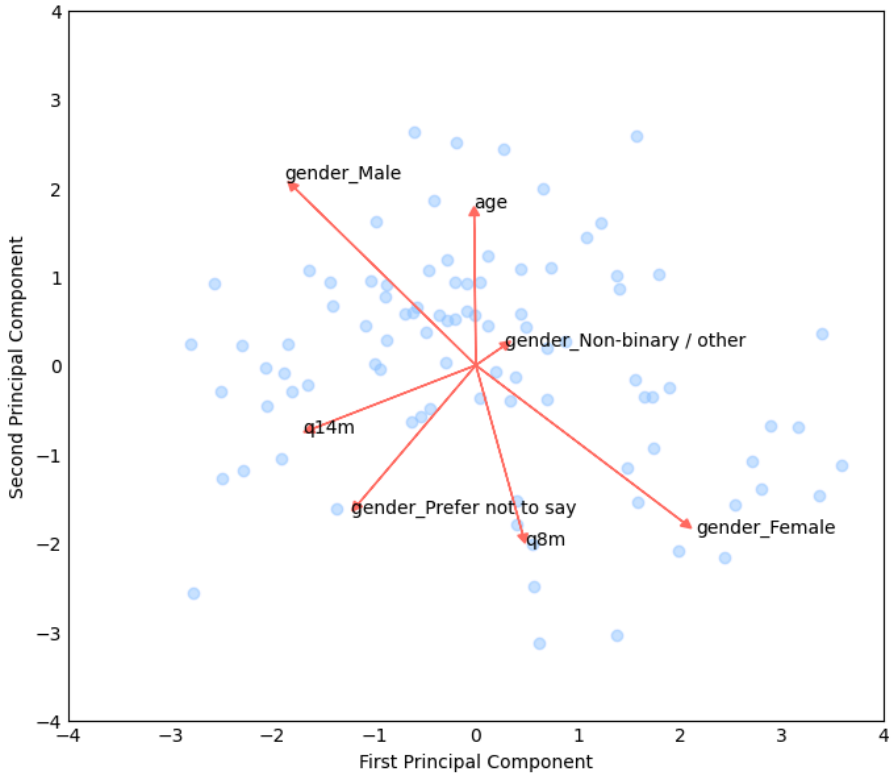

Figure 8. Dataset Spring 2025 Reduced to First Two Principal Components

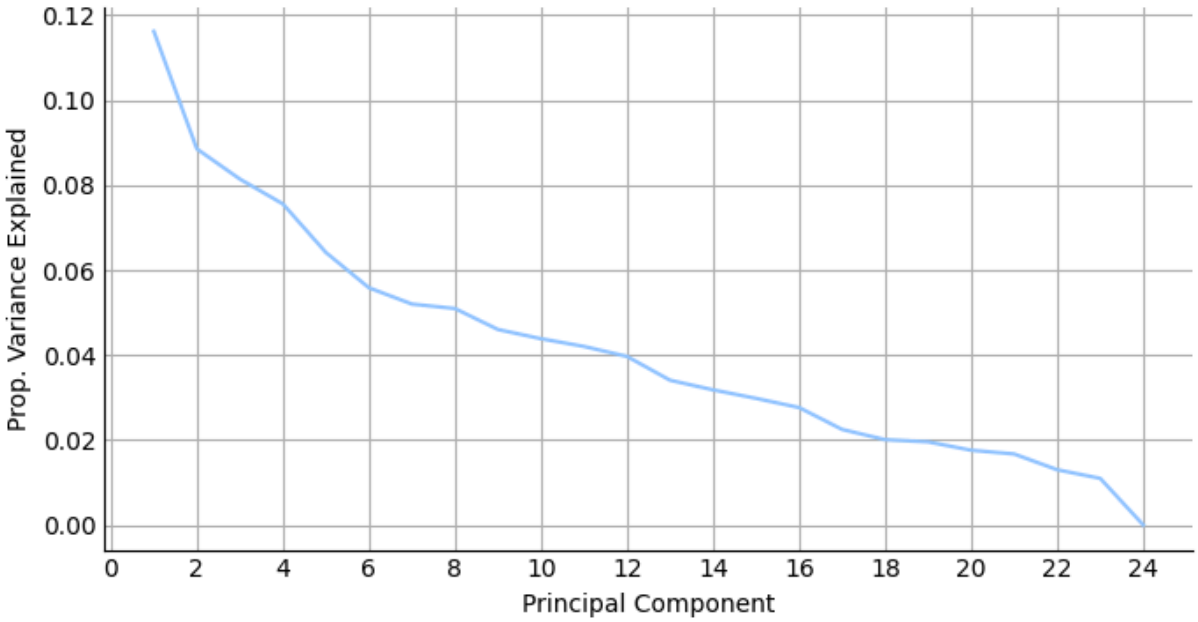

Figure 9. Scree Plot of PCA for Dataset Fall 2025

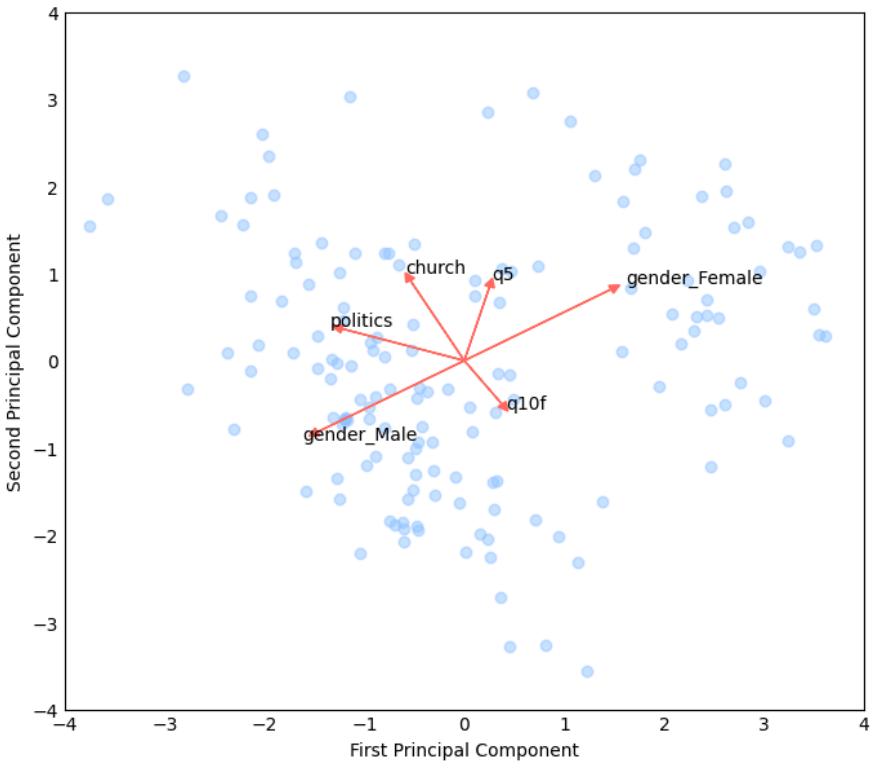

Figure 10. Dataset Fall 2025 Reduced to First Two Principal Components

It can be seen that the scree plots of each dataset do not have prominent elbow points, implying that responses to each question may not have much covariance with those of another question. It can also be seen that while vectors for male and female variables are necessarily pointed in opposite directions in the reduced dimensionality plots due to the nature of one-hot encoding, they are not always antiparallel. When there are "Non-binary / other" or "Prefer not to say" responses, this can cause the male and female vectors to become non-antiparallel. However, this does not always happen, as can be seen in Figures 2 and 8. This is not necessarily the case due to a lack of respondents that fall within non male or female categories, as the dataset "Fardina" (corresponding to Figures 3 and 4) had only one respondent identify as Non-binary/other, while the datasets for 2024 and Spring 2025 (corresponding to Figures 2 and 8) had multiple respondents identifying as non-binary/other or "prefer not to say".

Hypothesis Testing:

A total of 8 hypothesis tests were run. Thus, when using a standard significance level of 0.05, Bonferroni correction would adjust the significance level to 0.05/8 = 0.00625.

One question was whether or not the distribution of responses would be the same across surveys, given the variance in when they were administered as well as variations in which questions were asked and how. To isolate these variations from changing the gender of those in the "Am I the Jerk" scenario, the distributions of responses to two questions where pronouns were all gender-neutral (questions 5 and 12) were tested. Due to these responses being distributed across a non-neutral centered 3-point Likert scale and thus being unlikely to be normally distributed, a Kruskal-Wallis test was used to compare distributions instead of a two-way ANOVA. Despite the presence of a priming question in the dataset "Fardina," p-values only reached as low as approximately 0.113 and 0.112 for questions 5 and 12 respectively. Thus, since these values are higher than the adjusted alpha level, it cannot be concluded that there was significant variation in the judgement of respondents across surveys.

Regardless, in order to avoid introducing confounding variables, observations were not merged across datasets in further hypothesis tests.

In each dataset, most scenarios shared the gender of the individual(s) most affected by the potential jerk. Responses were then separated by whether the respondent shared the gender of the individual affected in the scenario – for example, if most scenarios had an individual being a potential jerk to females, then female responses were put in one group and all others (including male, non-binary/other, and "prefer not to say" responses) were put in another. A Mann-Whitney U test was used to compare groups rather than a t-test, since a t-test assumes continuous data. This process was then repeated for each dataset, resulting in a total of 5 additional hypothesis tests. P-values ranged from as high as 0.893 to no lower than 0.114. Below is a histogram showing the distributions of responses on the dataset that yielded the lowest p-value.

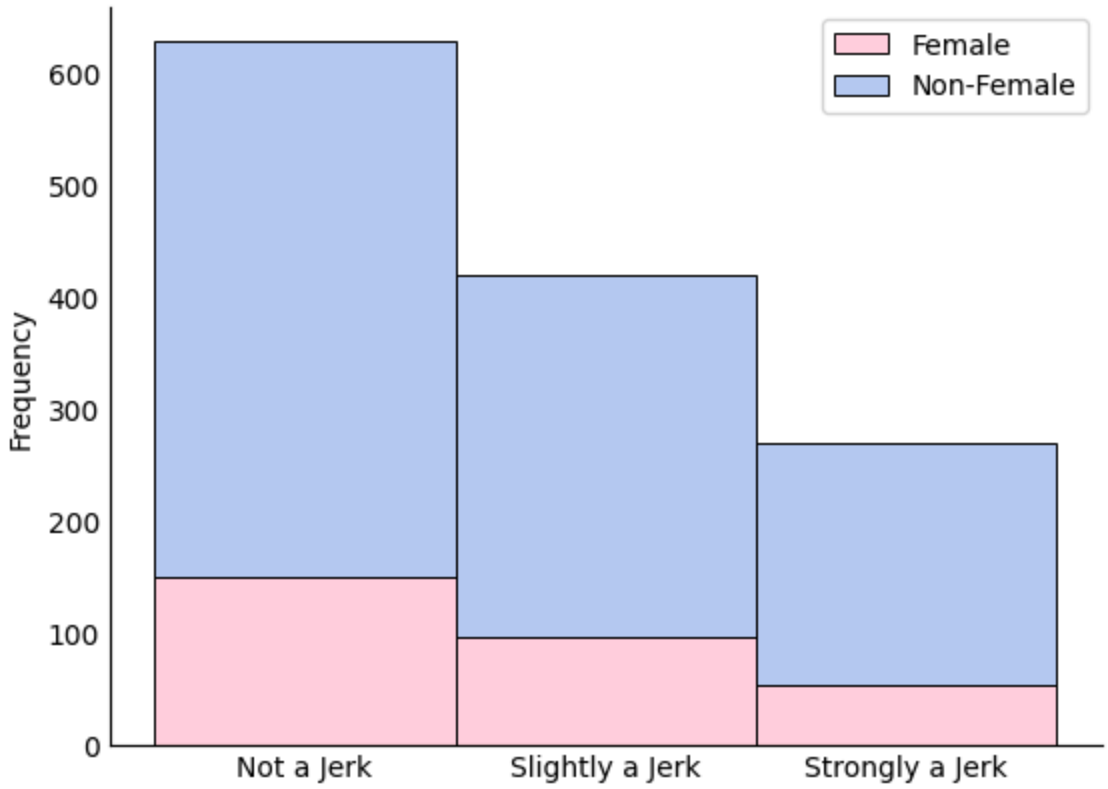

Figure 11. Distribution of "Am I the Jerk" Responses by Gender for Dataset Fall 2025

Again, since these values are higher than the adjusted alpha level, it cannot be concluded that there was significant variation in the judgement of respondents due to being biased towards one's own gender.

One final hypothesis test was run on the dataset "Fardina" to investigate differences between how genders self-reported their level of compassion. "Fardina" was the dataset that contained a priming question, asking for a yes or no answer to whether or not the respondent considered themselves a compassionate person. Responses were separated into responses from males and responses from females. There was one response that identified as non-binary/other; however, since there was only one response in this group, there were not enough observations to create a group for non-binary/other responses and test whether its distribution differed from the male and female groups. Thus, hypothesis testing was limited to comparing males and females. Once again, a Mann-Whitney U test was used to compare groups, and found a p-value of 0.565. Thus, we cannot conclude that there is a significant difference between the self-reported compassion levels between males and females.

Conclusions

These findings do not indicate that individuals judge morality differently based on whether they identify with a victim based on gender. Variation in responses to these surveys may not have been due to in-group bias based on gender, and instead may have been correlated with the metrics not considered within this report. Future analysis may investigate the other demographic metrics and see what effect, if any, they have on responses to “Am I the Jerk” scenarios. However, given that none of the PCA scree plots did not exhibit strong elbow points for any dataset, variation in responses may not be able to be explained with a single variable, as the total variance of the dataset is spread across many components.