Gender Bias in Science Fiction Fantasy

Lavanya Vijayan, Charis Chan, Joyce Ching, Inderpal Kaur

April 20, 2021

Abstract

Gender biases permeate both throughout the real world and literature. Especially in the context of the science-fiction and fantasy (SFF) genres that are heavily male dominated and have been found to influence public perceptions and acceptance of science, evaluating the relationship between gender and character perception can provide deeper insight into how gender influences STEM fields. In particular, based on a science fiction character's gender, are warmth and competence of the character and respect for them perceived differently?

We address this research question by recruiting MTurkers, university English departments, colleagues, friends, and family to answer a survey that asks participants to read a science fiction passage and evaluate the main character on the dimensions of warmth, competence, and respect. The intervention involved changing the passage to remove any identifiers and creating female and male versions of the passage by changing the pronouns associated with the main character. In the results, we found that those who received the male treatment passage perceived the main character to be warmer and more competent than those who received the female treatment passage. However, the small effect size and statistical insignificance of our results suggest that the quality of our original data may have impacted the observed patterns in our outcome measures.

Introduction

In the real world, humans and even machines sometimes report perceived differences between people of different genders that often align with gender stereotypes. These differences in manner and personality can also extend to perceptions about fictional characters in literature. Historically, the science-fiction and fantasy (SFF) genres in particular tend to be male-dominated in terms of authors, readership, and main characters (Flatt, 2018). This in conjunction with other factors has led some readers to observe that female main characters tend to be judged more harshly and along different criteria than their male counterparts in the genre (Konnikova, 2013). This raises the question: **Based on a science fiction character's gender, are warmth and competence of the character and respect for them perceived differently?** The answer to this question would be valuable to discover because the SFF genre in particular has been found to influence the perceptions and acceptance of science by the public (Menadue & Jacups, 2018). If there is a relationship between gender and character perception in SFF, this might be one lens through which we can investigate gender in science and STEM fields. Ultimately, literature has the power to either amplify stereotypes or encourage empathy and open-mindedness, and understanding how readers respond to characters can reveal potential dynamics of bias in the genre.

We are conducting a randomized controlled experiment to allow us to investigate this causal question and determine whether or not there is strong evidence to support the idea that a character's gender in SFF influences the perceptions of the readers.

To address the research question, the hypotheses are the following.

- Null hypothesis: There is no difference in perception between male and female versions of the character, in terms of warmth, competence, and respect.
- Alternative hypothesis: There is a difference in perception between male and female versions of the character, in terms of warmth, competence, and respect.

Based on existing gender stereotypes, our expectation is that, on average, female characters will be perceived as more warm and less competent and will be less respected than their male counterparts.

Related Works

Prior research in the area of gender stereotypes and perception has evaluated the perceptions that people have of others using the Stereotype Content Model (Fiske, 2018), or SCM. The SCM defines Warmth (trustworthiness, sociability) and Competence (capability, agency) as two common dimensions by which people tend to make judgements about individuals or groups (Fiske, 2018). Prior research using the SCM has shown that many stereotypes about social groups can be broken down along these dimensions to reveal certain associated emotional responses to those groups. For example, groups that typically rate high in perceived Warmth but low in perceived Competence (e.g. children, the elderly) often evoke emotions such as pity or sympathy, etc. In real life, researchers have observed differences between how men and women are perceived along these dimensions and how that can impact their social interactions. For example, one study found that perceived confidence in men in the workplace is correlated with their perceived Competence, whereas for women confidence and Competence are only correlated when they are also perceived as Warm (Mayo, 2016). Within gender categories, the Warmth-Competence scale has also shown distinctions between different "types" of men and women (Fiske, 2002). For example, "housewives" tend to receive high Warmth-low Competence scores, meanwhile "businesswomen" receive the opposite (low-high). For our experiment, we applied a similar evaluation scale to the context of SFF literature to examine gender bias.

Experiment

To address the research question, we carried out the experiment through a survey we created on Qualtrics. As part of the survey, each participant was given a short SFF passage to read followed by questions about their perceptions of the main character in the passage. We enabled a survey protection feature that prevented participants from taking the survey more than once.

Treatment

In an ideal setting, we would be able to gather responses to a novel or other real work of fiction by creating two versions with the same story except for the gender of the main character; however, recruiting participants to read a full-length novel would have taken too long and introduced the possibility that some respondents had already read the work or were aware of the original character's gender. Instead, we looked for an engaging passage (roughly 1000 words and about a 5 minute read) from a reputable author in the SFF genre. This passage was not the climax of the story since it would have been too easy for respondents familiar with the SFF genre to recognize the work. We used the introduction of *I, Robot* by Isaac Asimov because it fit all the criteria we were looking for: reputable author and story, short, and engaging. The original passage was an interview with the main character, Susan Calvin, about her life's work with robots and her retirement. We created two versions of the passage that signaled different gender identities for the main character through gender pronouns, and half of the participants of the study were randomly assigned to read the version of the passage that used typically masculine pronouns (he/his), while the other half read the version of the passage that used typically feminine pronouns (she/hers). The passage only referred to the main character through gender pronouns because we didn't want the character name to bring up unconscious name biases. We also replaced any notable details such as specific sci-fi phrases only Asimov uses, company names, and supporting character names that would likely lead someone to recognize the passage. Aside from the character's pronouns, the content of the passages were the same for the two different passages. We wanted to avoid priming participants to be more aware or sensitive to the character's gender before reading the passage, therefore we excluded explanations of the treatment and control conditions from the study description.

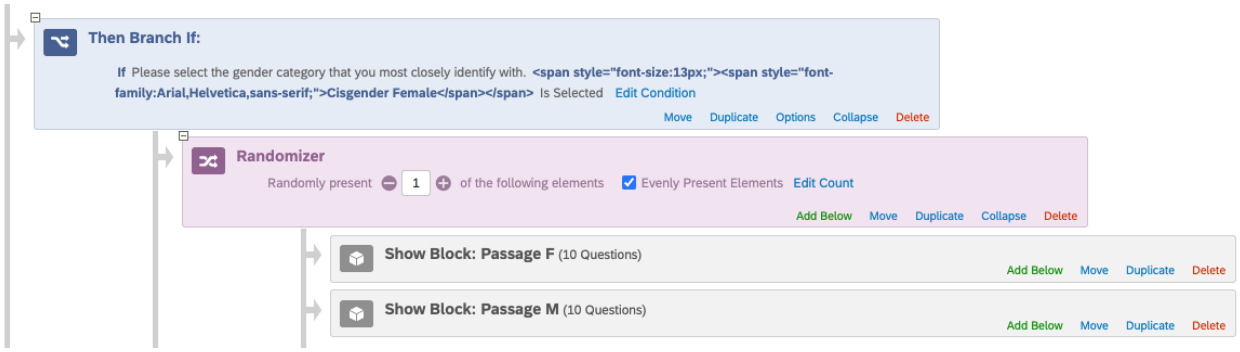


Figure 1: Conditional

An excerpt from the original piece and excerpts from the two treatment passages we constructed from it are displayed in Appendix 1.

Comparison Of Potential Outcomes

$R\ X_F\ O$

$R\ X_M\ O$

Our research design can be described using the above ROXO notation. This is a Posttest Only Randomized Experiment where we used a between subjects comparison to observe the differences between subjects who saw only one variation of the treatment. We began with using blocked random assignment (R) to assign participants to either the female passage treatment group or the male passage treatment group. Members of each group were then given the passage associated with their group (XF, XM) and evaluated for their perception of the main character's warmth, competence, and respectability after reading. As participants had no way of knowing about the main character beforehand, we could only evaluate them after the treatment was given.

Randomizaton Process

In the pilot survey, we used simple random assignment. Each participant had an equal probability of being assigned Treatment 1 (Passage with Female Character) and Treatment 2 (Passage with Male Character).

In the final survey, we used block random assignment, wherein we blocked by participant gender. Towards the beginning of the survey, the participant was presented with 8 gender categories and asked to select the one they most closely identify with. Based on their selection, they were assigned either Treatment 1 or Treatment 2, in such a way that within each gender category, assignments were evenly distributed between Treatment 1 and Treatment 2. This was specified by conditionals implemented in Survey Flow in our Qualtrics surveys — a conditional as shown below, for each of the following gender categories: *Cisgender Female*, *Cisgender Male*, *Transgender Female*, *Transgender Male*, *Transgender*, *Nonbinary*, *Gender Fluid*, *Other*.

The motivation for blocking by participant gender in the final study was because we anticipated heterogeneous treatment effects based on this variable. Specifically, we anticipated a possible relationship between one's gender and how one rates a character of a particular gender. For example, male participants might tend to rate the male character more highly than the female character; similarly, female participants might tend to rate the female character more highly than the male character. Thus, the treatment effect may vary considerably between participants of one gender category and participants of another gender category. If most participants of a particular category were assigned the same treatment, the outcomes of each treatment

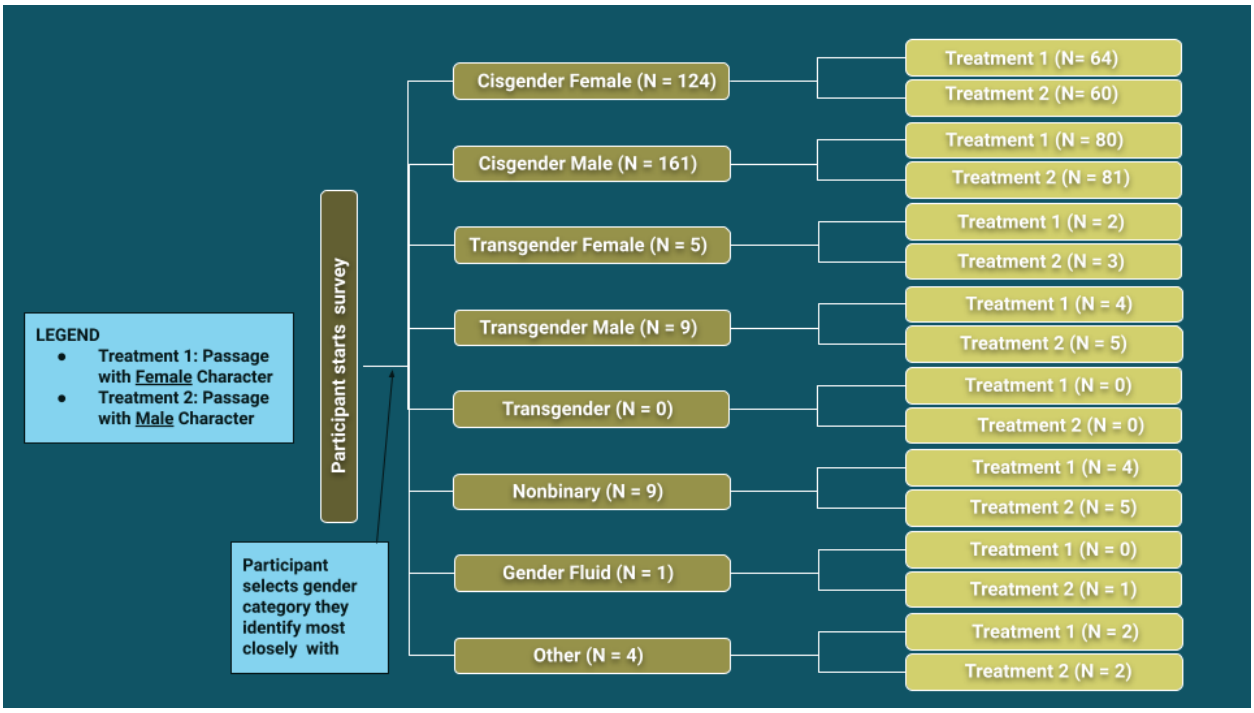


Figure 2: Consort Diagram

group would be polarized. Blocking by participant gender would help avoid that, as participants of each gender category would be evenly assigned to the treatments.

As displayed by the consort diagram above, the blocking worked, and we were able to achieve near-perfect even distribution of treatment assignment per participant gender category. For the categories where there is a difference of 1 between the number of participants of that category assigned to Treatment 1 and the number of participants of that category assigned to Treatment 2, that is due to the number of participants of that category being odd. For the categories where the difference is greater than 1, that is due to attrition.

Outcome Measures

To track whether readers make similar distinctions in judgment when it comes to fictional characters, we adapted the Warmth-Competence measures from the Stereotype Content Model to assess how character gender impacts reader perceptions.

After reading the passage, participants were asked a series of questions to gauge their perception of the main character. These questions were of the form: In your opinion, how _____ is the interviewee in the passage? The blanks were replaced by either Warmth terms (warm and well-intentioned) or Competence terms (competent and capable). We included an additional question asking, "How much do you respect the interviewee in the passage?" to consider an additional dimension of the character and any insights it may provide from if correlated with the other variables. As shown in Figure 3, participants' responses were recorded on a 5-point Likert scale, where 1 is "not at all" and 5 is "extremely" (Fiske et al., 2002).

The variables Warmth and Competence were calculated by taking the average responses of the associated questions. For example, someone who answered "3" for "In your opinion, how warm is the interviewee in the passage?" and "5" for "In your opinion, how well-intentioned is the interviewee in the passage?" would have a "Warmth" value of 4. Taking the average of multiple variables to calculate Warmth and Competence allowed us to treat Warmth and Competence as continuous variables when running the following tests and

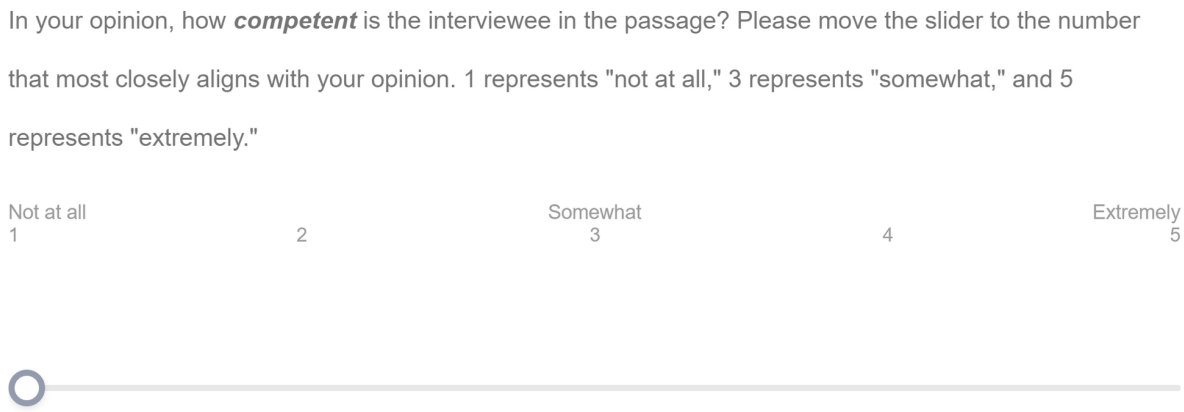


Figure 3: Question Format

regressions instead of as ordinal variables. The Respect variable was directly taken from responses to the question "How much do you respect the interviewee in the passage?" and remained as an ordinal variable. In total, we had three metrics by which we evaluated participants: Warmth, Competence, and Respect.

Covariates

In addition to being randomly assigned a version of the passage to read, participants were evenly blocked according to their gender to reduce the potential for the reader's gender to introduce major variations in character perception. Further covariates that we collected included basic information about the participant (i.e. gender, age, education) and the participant's reading habits (i.e. How much do you enjoy reading compared to doing other things? How much do you like science fiction and fantasy compared to other things you read?). We used the information about reading habits to investigate potential correlations between past experience with the genre and character perception; if there are differences between how experienced and inexperienced SFF readers respond, this could be an indication that the two groups have some fundamental differences in how they understand gender in SFF due to their reading habits.

To assess whether participants were actually responding to passage, we also added comprehension/recall questions that demonstrated the participant's understanding of the reading immediately after they were shown to passage. Specifically, we asked them the following (in the final study):

- What does the interviewer think the interviewee never does?
- What does the interviewee suggest the interviewer has not experienced?
- What is the interview being conducted for?

The correct answers to these questions were explicitly stated in the passage and should have been clear to respondents who read the text comprehensively according to the survey instructions. The effectiveness of these comprehension questions was confirmed by the high percentage of Slack/e-mail respondents who passed the attention check questions. Additionally, we also asked them "Did the passage look familiar to you?" to see if the familiarity of the passage or having read the full text was correlated with responses that differed from people who were reading the passage for the first time, as well as the effectiveness of our efforts in ambiguating the passage.

Pilot Study

Pilot Study Participants

Participants of the pilot study were recruited through reaching out to friends and family and posting in our 5th year MIDS cohort Slack channel. Participants of the pilot study were mainly friends, family, and cohort members. There were 27 participants in total.

Pilot Data

The distribution of each outcome measure in the pilot data is displayed in Appendix 2.

Power Calculation

Doing a pilot study was a way to see if our study was reasonable and if there were any changes we needed to make before we released our final study. We conducted power calculations on the pilot study data to predict the appropriate sample size needed in our final study. We calculated for a power of 0.8 for all three outcome variables and got varying results for what sample size was necessary. We set a power of 0.8 because it was the standard in research and we thought an 80% probability of detecting an effect, given that the effect is really there was a high enough probability. Our final study will conduct t-tests of our outcome variables since we would have a sample larger than 30 respondents so we also used the power calculations on a t-test to be consistent with our final study. We computed the difference in means and pooled standard deviation of our metrics: warmth, competence, and respect to input into the power calculation. We needed 1550 respondents in total (775 respondents for each treatment) to have a power of 0.8 for our Warmth metric that had a difference in mean of 0.08. With an even smaller difference in means (0.02) for our Competence variable we would have needed 46,464 total respondents. Both of these sample sizes were unreasonable in the time and resources we had available to gather responses. However, our Respect variable had a difference in means of 0.6 and with a power of 0.8 we would need only 100 total respondents (50 respondents in each treatment). This result gave us hope that if we could get at least 100 respondents we would have enough power to be confident in our results if we end up rejecting the null hypothesis in the final study.

Changes From Pilot Study To Final Study

A handful of changes were made to the survey after the pilot study. First, we shifted to blocked random assignment by participant gender instead of simple random assignment. The motivation for this was discussed earlier in the randomization section of the report. Additionally, we added more instructions at the beginning of the survey — participants were told to do their best to complete the survey in one sitting and that the survey would take approximately 10 minutes. This was done to motivate them to take the survey since it should not take more than 10 minutes and to encourage them to complete the survey sooner rather than later.

We also added a question asking for education level, which we thought would be a relevant covariate, as it would help inform us on participants' reading levels, which can impact how they perceive characters. Moreover, we addressed the ambiguity in the third post-treatment attention check question, which at the time had read "What has the interviewer not experienced" and rephrased it to "What does the interviewee suggest the interviewer has not experienced?" We underlined the word "interviewee" to explicitly set it apart visually from the word "interviewer" in every place where both words were in the same question. Furthermore, we lowered the number of page breaks from 4 to 2 in order to keep participants more engaged and feel less intimidated by the number of steps to complete the survey.

Lastly, we disabled the back button so that participants cannot change their answers back and forth and that they respond to the post-treatment questions from their memory of the passage they read.

Final Study

Final Study Participants

There were 313 participants total in the final study. We aimed to collect as many responses as possible. We created two Qualtrics surveys, one for distributing to MTurk and the other for distributing to UC Berkeley School of Information Slack channels and University English Departments across California. The only difference between the two is the generation of a completion code that is required for MTurk, as a way to confirm which MTurk workers have completed the survey and thus are eligible for pay. Our budget of $500 informed how many responses we could pay to get from MTurk, so we requested the maximum number of responses we could afford. We received 240 responses from MTurk, all of which were complete.

We received 106 responses from UC Berkeley School of Information Slack members and University English Department members collectively, 73 complete responses and 33 incomplete responses. All of the participants who gave incomplete responses did not answer the post-treatment questions, so they chose to leave the survey either right after answering the pre-treatment (covariate) questions or right after seeing their assigned passage.

To recruit participants from Slack, we posted to the School of Information program-specific channels (#mids-announcements, #mims-announcements, #mics-announcements), as well as a general channel that anyone in that Slack workspace could be in (#noise).

To recruit participants from University English Departments across California, we scraped email addresses of department chairs and advisors at 144 universities and we schedule sent emails to half of them on Tue, Mar 16, 7:48 AM and to the other half of them on Wed, Mar 17, 7:48 AM. We emailed them the survey information and asked them to forward it to faculty and students in their department.

We reached out to peers on Slack and University English Departments after getting the maximum responses from MTurk, in order to achieve a larger sample size and recruit people who we believed would be more likely to be invested in participating and willing to read a passage without pay.

The messages we wrote in Slack and over email are documented in Appendix 3.

Final Data

The distribution of each outcome measure and distribution of each covariate in the final data are displayed in Appendix 4.

Exploratory visualizations of trends between covariates and the source (source being whether a response is from MTurk or Slack/Email) are displayed in Appendix 5.

Initial Results

To analyze our results, we took the average Warmth rating and Competence rating for each participant from the Warmth-specific and Competence-specific questions respectively. We also took the Respect rating directly from the Respect-specific question. We compared these average ratings between the group that read the passage with a male main character and the group that read the passage with a female main character. We tested whether there were significant differences between the average measures for the two groups that can be attributed to the character's gender.

Difference In Means

Difference In Means For Warmth

The average `warmth` rating given to the female character was 3.42 and the average `warmth` rating given to the male character was 3.417. The average treatment effect with the `warmth` metric (the average difference in perceived `warmth` between female character and male character) is 0.003. This test resulted in a p-value of 0.977 and standard error of 0.094.

Difference In Means For Competence

The average `competence` rating given to the female character was 3.926 and the average `competence` rating given to the male character was 3.924. The average treatment effect with the `competence` metric (the average difference in perceived `competence` between female character and male character) is 0.003. This test resulted in a p-value of 0.972 and standard error of 0.077.

Difference In Means For Respect

The average `respect` rating given to the female character was 3.795 and the average `respect` rating given to the male character was 3.783. The average treatment effect with the `respect` metric (the average difference in perceived `respect` between female character and male character) is 0.011. This test resulted in a p-value of 0.911 and standard error of 0.103.

The effect size for `respect` was largest, which indicates that the difference between how much respect participants have for the female character and how much respect participants have for the male character was largest, compared to the differences in the other metrics of perception (`warmth` and `competence`).

The effect direction is positive for all three metrics (`warmth`, and `competence`, and `respect`), which indicates that on average, the female character was given higher ratings than the male character, along the three metrics. This does not completely align with our expectations — while we did expect the female character's warmth rating to be higher than the male character's on average, we did not expect the female character's competence and respect ratings to also be higher than the male character's on average.

Linear Regression

Before examining our regression results, we performed a covariate balance check to test for inconsistencies in our randomization scheme. To compare a null intercept-only model to a full model with the covariate information we collected, we conducted an F-test to analyze the amount of additional variance explained by the full model. With a p-value of 0.257, we failed to reject the null hypothesis. In terms of our balance check, the test does not provide strong evidence of covariate imbalance or randomization errors across the treatment groups.

As seen in the table below, the regression results for each of the outcome measures demonstrate a small negative treatment effect. The model for Warmth showed the treatment effect of receiving the Male Passage was a decrease of -0.019 points on average for the Warmth rating. This supported our initial hypothesis, which assumed based on common gender stereotypes that the female character would be perceived as warmer than her male counterpart. In terms of practical significance however, a difference of -0.019 on a 5-point Warmth scale is not a very substantive gap.

Similarly, the regression model for Competence showed the treatment effect of receiving the Male Passage was a decrease of -0.068 points on average for the Competence rating. This contradicted our initial hypothesis, which assumed that the female character would be perceived as less competent than her male counterpart. In terms of practical significance, the estimated Competence treatment effect was roughly 3 times larger than the estimate for Warmth, which could be an indication that the passage context may be more impactful on

Table 1: Full Data Regression

	Dependent variable:		
	warmth	competence	respect
	(1)	(2)	(3)
Treatment (male passage)	−0.019	−0.068	−0.039
	(0.078)	(0.073)	(0.095)
Block: Cisgender Male	0.010	−0.018	−0.016
	(0.080)	(0.076)	(0.103)
Block: Gender Fluid	−0.383***	0.308***	0.268*
	(0.109)	(0.106)	(0.138)
Block: Nonbinary	−0.643**	−0.462*	−0.410
	(0.313)	(0.236)	(0.303)
Block: Other	−0.156	−0.373	−0.634
	(0.363)	(0.397)	(0.691)
Block: Transgender Female	0.170	0.180	−0.092
	(0.396)	(0.214)	(0.294)
Block: Transgender Male	0.0001	−0.239	−0.146
	(0.240)	(0.224)	(0.209)
Age	−0.004	0.007*	−0.001
	(0.005)	(0.004)	(0.006)
Bachelor's degree	0.681***	−0.494**	0.179
	(0.217)	(0.220)	(0.243)
Doctorate degree	0.228	−0.254	0.269
	(0.369)	(0.463)	(0.436)
High school/GED	−0.125	0.104	−0.321
	(0.256)	(0.250)	(0.322)
Master's degree	0.769***	−0.491**	0.209
	(0.222)	(0.220)	(0.250)
Enjoy reading	0.080*	0.059	−0.013
	(0.047)	(0.052)	(0.066)
Books read	0.007	0.004	0.006
	(0.004)	(0.004)	(0.005)
Enjoy sci-fi	0.246***	0.154***	0.250***
	(0.041)	(0.046)	(0.060)
Constant	1.704***	3.338***	2.843***
	(0.282)	(0.321)	(0.394)
Observations	313	313	313
R^2	0.346	0.128	0.152
Adjusted R^2	0.313	0.084	0.109
Residual Std. Error (df = 297)	0.685	0.648	0.855
F Statistic (df = 15; 297)	10.488***	2.912***	3.547***
Note:			*p<0.1; **p<0.05; ***p<0.01

perceptions of Competence than Warmth. Nevertheless, a treatment effect of -0.068 points on a 5-point scale is not very substantial.

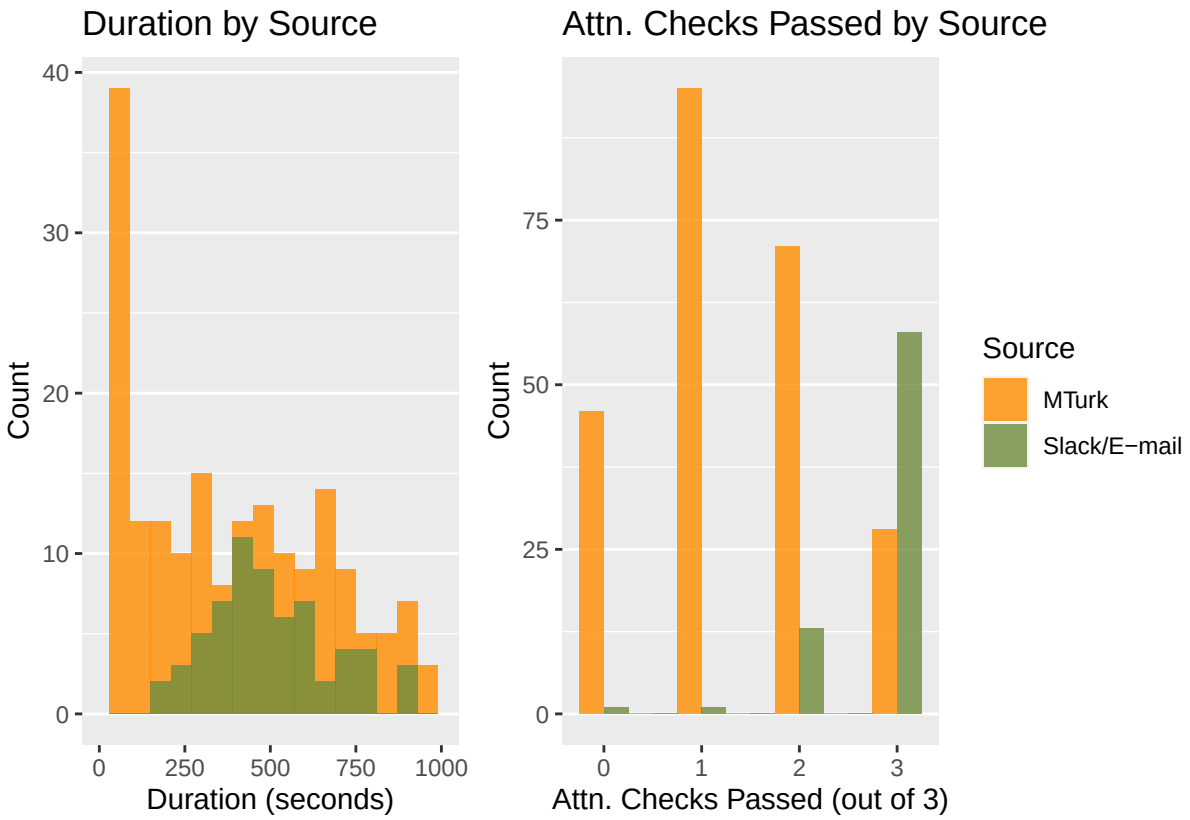
Finally, the model for Respect showed the treatment effect of receiving the Male Passage was a decrease of -0.039 points on average for the Respect rating. Again, this was not in favor of our initial hypothesis, which assumed based on common gender stereotypes that the female character would be perceived as less respectable than her male counterpart. The estimated effect for Respect was roughly 2 times larger than the estimate for Warmth but smaller in size than the effect for Competence. However, this treatment effect is still insubstantial on a 5-point scale.

Overall, all three models show treatment effects that are not substantially or significantly different from 0 at the 95% confidence level. Even so, looking at our best estimates shows that out of the three measures, the estimates for Competence and Respect are the largest in size. This might indicate that the gender of the character for this particular passage made more of an impact along these two dimensions than along Warmth. The focus on Competence especially seems reasonable considering the context of the passage, which is about the character's specific achievements over the course of their career. The direction of these two larger effects however was surprising as we had expected that the female character would receive lower ratings for Competence and Respect. Although our models do not reveal why we found the estimated treatment effects that we did, there are several possibilities that might be interesting to explore in future work, including running another study with a larger sample size/more power, selecting a less recognizable passage, or choosing a passage in which the character was originally male rather than female.

Exploration

This section of the report details exploratory analysis we performed on a subset of the data, in order to investigate trends on the subset. We acknowledge these trends may not be reflected by the full data.

Exploratory Filtration Of Data



As the visualization above shows, participants from MTurk tended to take less time to complete the survey and pass fewer attention checks, compared to participants from Slack/Email.

passage_gender	source	avg_warmth	avg_competence	avg_respect	recognized
PassageF	mturk	3.643	3.899	3.866	0.765
PassageF	slack_email	2.703	4.014	3.568	0.108
PassageM	mturk	3.653	3.884	3.926	0.719
PassageM	slack_email	2.625	4.056	3.306	0.111

As the table above shows, the proportion of participants who recognized the passage from MTurk is about 0.70 higher than the corresponding proportion from Slack/Email. Moreover, participants from MTurk tended to give ratings that were in the middle of the Likert scale on average and their average ratings for `warmth`, `competence`, and `respect` were relatively close together. There was more distinction between ratings for the three metrics given by participants from Slack/Email.

This leads us to believe that responses from MTurk might not be as thorough as responses from Slack/Email.

Thus, we filter the data, creating a subset that only includes responses coming from Slack/Email.

Results From Filtered Data

Difference In Means On Filtered Data

Difference In Means For Warmth On Filtered Data

The average `warmth` rating given to the female character was 2.703 and the average `warmth` rating given to the male character was 2.625. The average treatment effect with the `warmth` metric (the average difference in perceived `warmth` between female character and male character) is 0.078. This test resulted in a p-value of 0.665 and standard error of 0.182.

Difference In Means For Competence On Filtered Data

The average `competence` rating given to the female character was 4.014 and the average `competence` rating given to the male character was 4.056. The average treatment effect with the `competence` metric (the average difference in perceived `competence` between female character and male character) is -0.042. This test resulted in a p-value of 0.834 and standard error of 0.203.

Difference In Means For Respect On Filtered Data

The average `respect` rating given to the female character was 3.568 and the average `respect` rating given to the male character was 3.306. The average treatment effect with the `respect` metric (the average difference in perceived `respect` between female character and male character) is 0.262. This test resulted in a p-value of 0.337 and standard error of 0.276.

From performing exploratory difference in means on the filtered data, we observe that the effect direction has changed for `competence` from positive to negative. This indicates that among the participants coming from Slack or Email, participants perceived the female character's competence to be lower than the male character on average. The female character being perceived as having less competence aligns with our expectation.

We also observe that the effect size on the filtered data was larger across all three metrics, compared to the unfiltered data. This indicates that the difference in perception of `warmth`, `competence`, and `respect` between female character and male character is greater among participants who came from Slack than the all participants as a whole.

The exploratory difference in means tests on the filtered data were not statistically significant, but the p-values were smaller than the respective tests on the unfiltered data.

Linear Regression On Filtered Data

We applied the same regression tests as before to the potentially less noisy/higher quality subset of data from the Slack/E-mail sources. The filtered model for Warmth showed the treatment effect of receiving the Male Passage was an increase of 0.011 points on average for the Warmth rating. This result shows a reversal in the direction of the effect we observed in the full dataset and now contradicts our initial hypothesis that the female character would be perceived as warmer than her male counterpart. Although the effect is reversed in direction, the size is comparable to the prior estimate which means that it is still not a practically substantive effect on a 5-point Warmth scale.

Similarly, the filtered model for Competence showed the treatment effect of receiving the Male Passage was an increase of 0.051 points on average for the Warmth rating. This result also shows a reversal in the direction of the effect we observed in the full dataset and now supports our initial hypothesis that the female character would be perceived as less competent than her male counterpart. Again, the estimated Competence treatment effect was roughly 5 times larger than the estimate for Warmth, which maintains the pattern in effect size we noticed in our earlier analysis.

Table 3: Filtered Data Results

	Dependent variable:		
	warmth	competence	respect
	(1)	(2)	(3)
Treatment (male passage)	0.011	0.051	−0.088
	(0.177)	(0.184)	(0.290)
Block: Cisgender Male	0.190	−0.032	−0.478
	(0.192)	(0.229)	(0.324)
Block: Nonbinary	−0.381	−0.468*	−0.485
	(0.249)	(0.261)	(0.357)
Block: Other	−2.092***	−2.316***	−3.533***
	(0.568)	(0.406)	(0.528)
Block: Transgender Female	−0.097	0.673***	−0.724**
	(0.262)	(0.220)	(0.346)
Block: Transgender Male	0.525**	−1.186***	−0.393
	(0.266)	(0.269)	(0.466)
Age	0.007	0.013**	−0.012
	(0.010)	(0.005)	(0.013)
Bachelor's degree	0.362	−0.638***	−0.101
	(0.264)	(0.219)	(0.373)
Doctorate degree	0.573	−0.415	0.752
	(0.604)	(0.395)	(0.505)
High school/GED	0.107	−0.323	−0.617
	(0.279)	(0.198)	(0.380)
Master's degree	0.152	−1.141***	0.043
	(0.343)	(0.304)	(0.478)
Enjoy reading	−0.157**	−0.120	−0.248**
	(0.067)	(0.085)	(0.110)
Books read	0.004	0.002	−0.0005
	(0.005)	(0.005)	(0.006)
Enjoy sci-fi	0.037	0.051	0.123
	(0.055)	(0.066)	(0.100)
Constant	2.702***	4.513***	4.935***
	(0.396)	(0.378)	(0.666)
Observations	73	73	73
R^2	0.260	0.303	0.229
Adjusted R^2	0.082	0.135	0.043
Residual Std. Error (df = 58)	0.723	0.788	1.130
F Statistic (df = 14; 58)	1.459	1.799*	1.229
Note:			*p<0.1; **p<0.05; ***p<0.01

Finally, the filtered regression for Respect showed the treatment effect of receiving the Male Passage was a decrease of -0.088 points on average for the Respect rating. Unlike the previous two filtered models, this result maintains the direction of our initial estimate and shows an effect size that is 2 times larger than the original.

Overall in the exploratory regression analysis, all three models show treatment effects that are not substantially or significantly different from 0 at the 95% confidence level. Additionally, with a smaller sample size, the standard errors for these estimates are larger. Even so, we note that an interesting element of our estimates is that the directions changed for Warmth and Competence but not for Respect when using the filtered subset of data. This may be an indication that more higher quality data from participants that pass the attention checks would reveal underlying patterns that the original MTurk data obscured with "noisy" or middle-of-the-road responses. In particular, we would be interested in further examining the direction of the effects for Warmth and Respect, which contradicted our expectation.

Conclusion

Although the results from our final study initial results showed negative effects along the warmth, competence, and respect metrics from reading the male passage compared to the female passage, we found that our experiment was underpowered and presented low practical significance and no statistical significance due to the small effect size and large p-value. Upon further examination of our survey responses, we found that our different methods of sourcing participants had resulted in a range of response qualities, specifically that participants sourced through Slack and e-mail passed a much larger proportion of our passage attention checks. In response to this finding, we subset our data to only include participants sourced through Slack and e-mail. After performing tests on this subset of data, we found that the male character was perceived as being warmer and more competent, but less respected. This finding aligned with our expectation along the dimension of competence, and did not align with our expectation along the dimensions of warmth and respect. It indicates that the quality of our original data may have impacted the patterns we observed for our outcome measures and that the question of gender stereotypes in science fiction requires more investigation in the context of different passages to resolve.

One of the limitations of our study is that it is highly dependent on the passage chosen for the treatment. From this study alone, it may be difficult to extrapolate the results because it is possible that something specific to the passage itself was what prompted different character perceptions (e.g. if the passage happened to describe an activity that is stereotypically seen as masculine, maybe the female main character might be perceived differently than the male main character specifically in this short context, but not in general across SFF literature). In addition, the SFF genre comes with its own stereotypes about gender, especially in terms of science and technology, which means that any differences in perception may be a function of the genre and not applicable to other works outside of SFF. Lastly, character perceptions may also be influenced by general enjoyment of the passage.

In future studies, we would be interested in seeing the treatment effects for a less recognizable passage and where the female character is presented as less of a "cold scientist". We would also like to find ways to obtain a larger, higher quality sample of participants where they read the passages closely enough to answer the attention check questions accurately. However, this will be challenging to achieve as it is difficult to motivate participants and anticipate quality of responses prior to applying the treatment. Targeting people who would be willing to read comprehensively will limit the population we can extrapolate our findings to and may potentially bias our results to a specific subset of the population. Shortening the passage also gives readers less context through which to evaluate the characters and makes whichever passage we choose less explicitly of the SFF genre. Moreover, a shorter passage might motivate more participants to complete the survey, leading to less attrition and larger sample size, as well as enable them to read the story more closely without skipping or skimming over details, leading their responses to the post-treatment questions to be more representative of their perceptions.

To further expand our research, we would also be interested in applying our current procedure to different

genres other than SFF and expanding the range of character genders. Our current design tests specifically for differences in perception of male and female characters in SFF, but it would be relevant to investigate whether the results of the study apply beyond the context of the specific passage and genre as well as how perceptions may change for non-binary or gender non-conforming characters.

Appendix

Appendix 1. Process of constructing treatments

LEGEND

Blue box := indicates character name and/or gender

Green underline := indicates passage source

We transformed the short story into two treatments by:

- Replacing indicators of literature source with less revealing words
- Replacing identifiers of gender with gender pronouns corresponding to each treatment

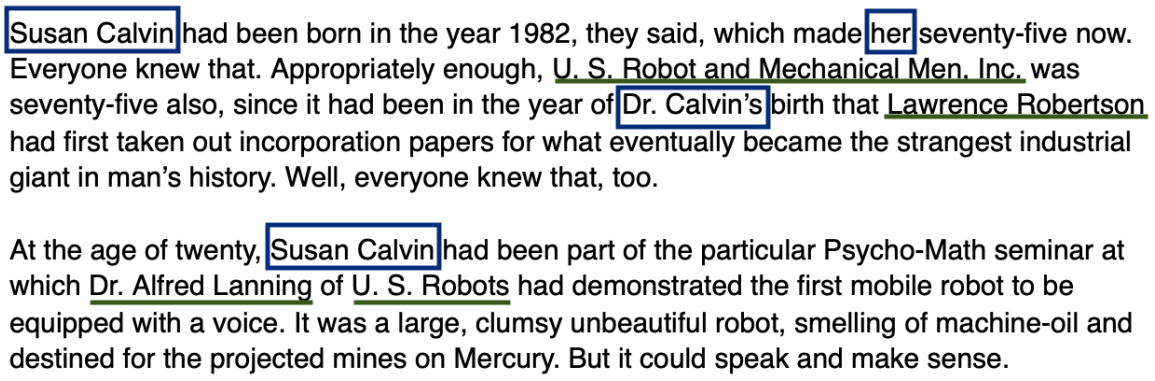


Figure 4: Original excerpt from the short story "I, Robot" by Isaac Asimov

She had been born in the year 1982, they said, which made her seventy-five now. Everyone knew that. Appropriately enough, Robot, Inc. was seventy-five also, since it had been in the year of her birth that the CEO had first taken out incorporation papers for what eventually became the strangest industrial giant in human history. Well, everyone knew that, too.

At the age of twenty, she had been part of the particular Psycho-Math seminar at which the director of research of Robot, Inc. had demonstrated the first mobile robot to be equipped with a voice. It was a large, clumsy unbeautiful robot, smelling of machine-oil and destined for the projected mines on Mercury. But it could speak and make sense.

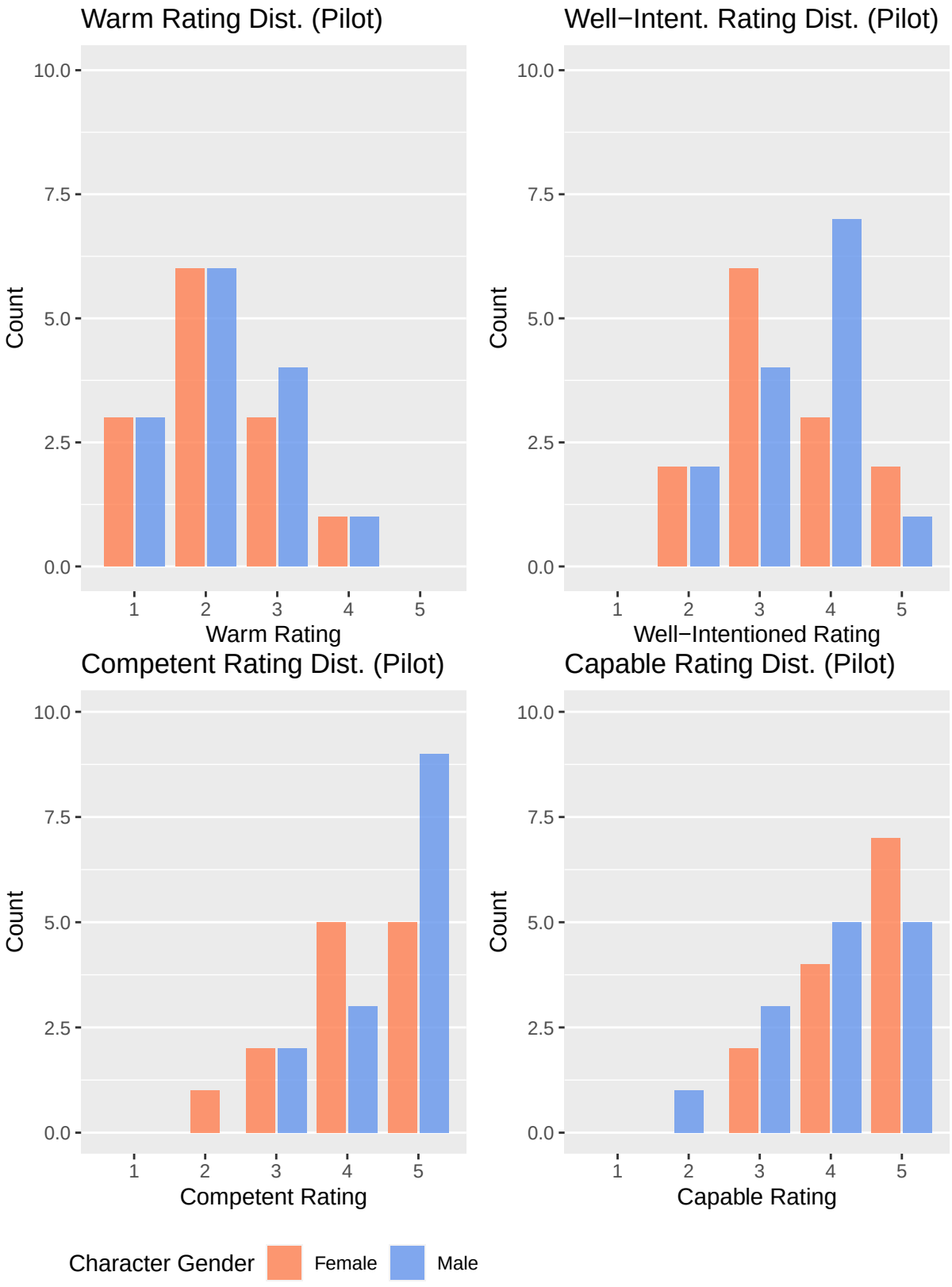
Figure 5: Modified excerpt from Treatment 1: Passage with Female Character

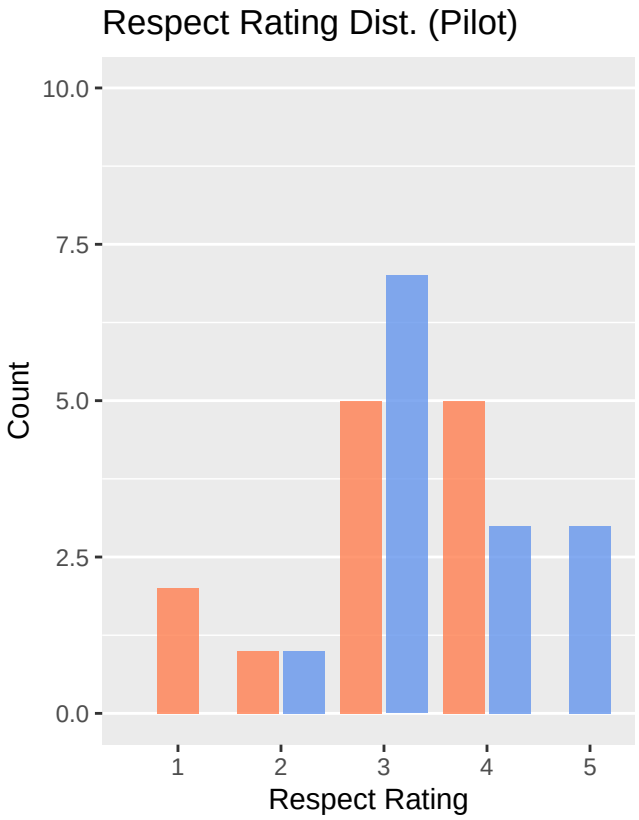
He had been born in the year 1982, they said, which made him seventy-five now. Everyone knew that. Appropriately enough, Robot, Inc. was seventy-five also, since it had been in the year of his birth that the CEO had first taken out incorporation papers for what eventually became the strangest industrial giant in human history. Well, everyone knew that, too.

At the age of twenty, he had been part of the particular Psycho-Math seminar at which the director of research of Robot, Inc. had demonstrated the first mobile robot to be equipped with a voice. It was a large, clumsy unbeautiful robot, smelling of machine-oil and destined for the projected mines on Mercury. But it could speak and make sense.

Figure 6: Modified excerpt from Treatment 2: Passage with Male Character

Appendix 2: Distributions of outcome measures in pilot data




Respect Rating Dist. (Pilot)
10.0
7.5
5.0
2.5
0.0
Count
1
2
3
4
5
Respect Rating

Appendix 3: Verbiage for recruiting participants via Slack and Email

Message sent to Slack Channels

Hi y'all! I'm a 5th year MIDS student; my team and I are currently conducting an experiment that requires participants to read a short science fiction passage and answer some questions about it. It would be a tremendous help if you could participate. The survey should only take a few minutes to complete and will hopefully be a fun read! Here is the survey link: [SURVEY LINK]. Thanks for your time!

Email sent to University English Departments

Subject: Science-Fiction Survey: Please forward to your department and students

Body: Hello [Staff Name],

My name is [Team Member Name] and I am currently a student in UC Berkeley's Masters of Information and Data Science program. My team and I are currently conducting an experiment that requires participants to read a short science-fiction passage and answer a few questions about it. It would be a tremendous help if you could forward our survey to your department staff and students. The survey should only take a few minutes to complete and will hopefully be a fun read!

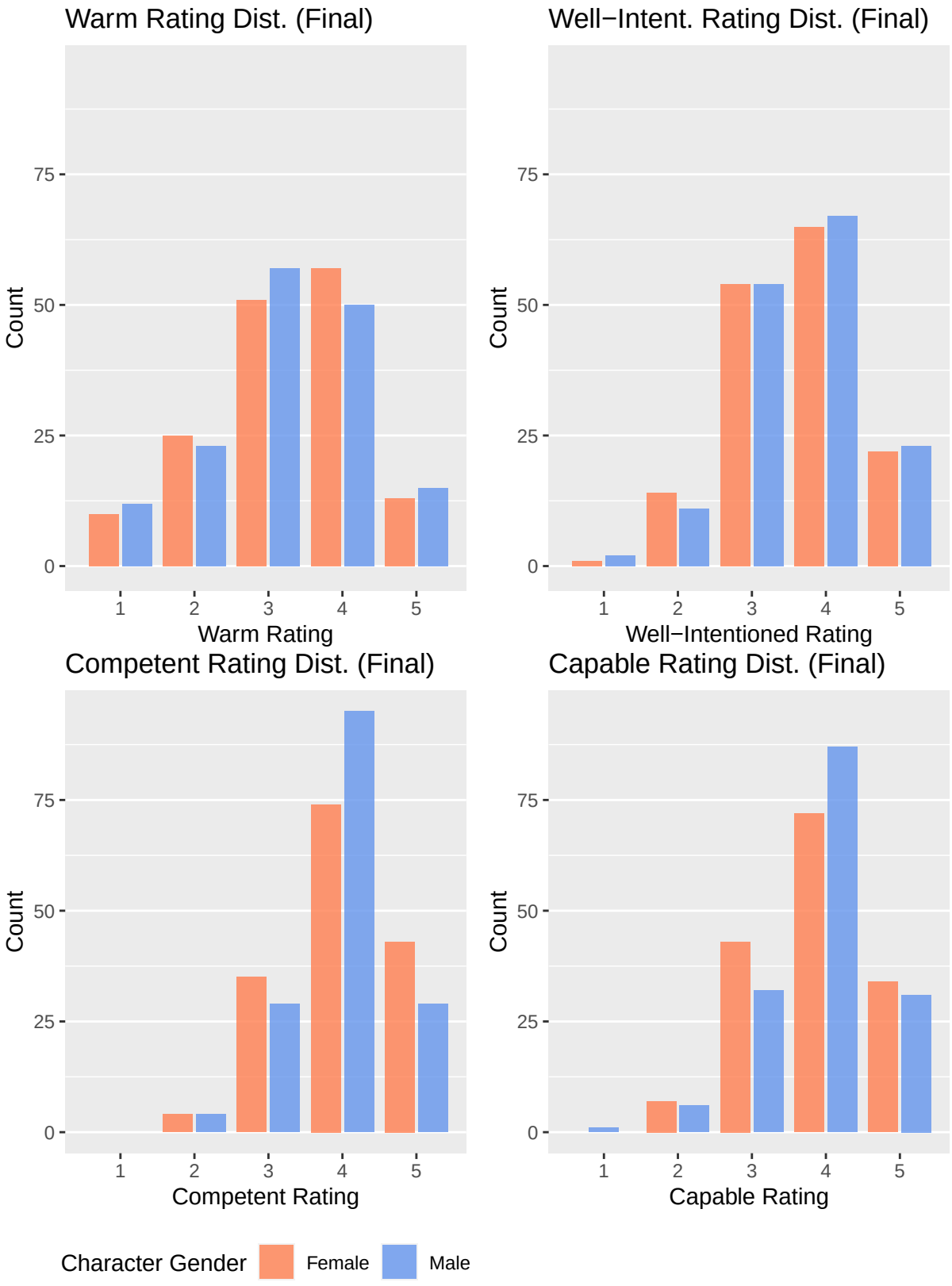
Here is the survey link: [SURVEY LINK]

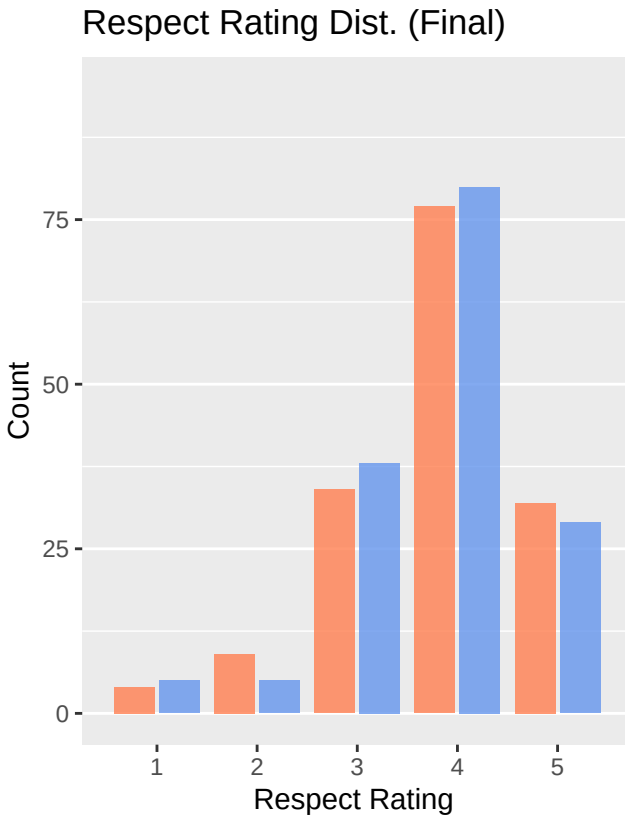
Please let me know if you have any questions or concerns.

Thank you for your time and best regards, [Team Member Name]

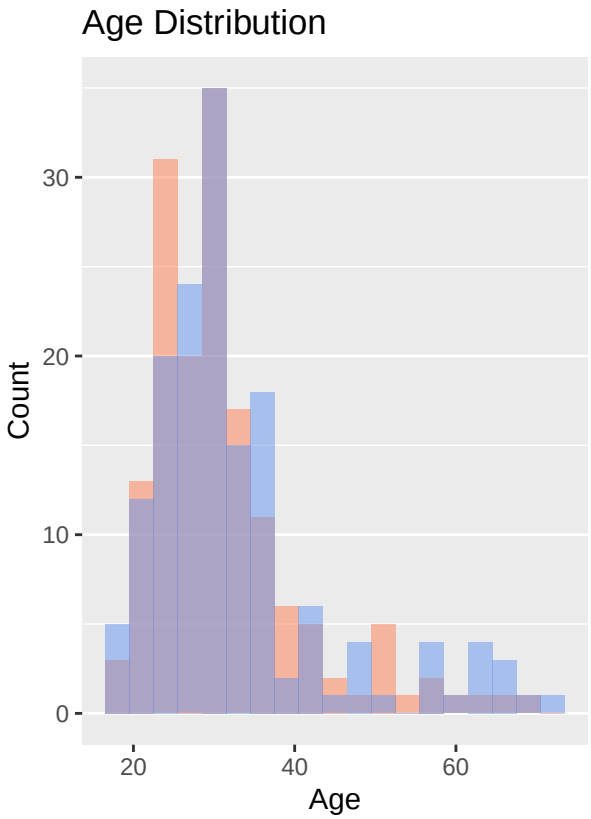
Appendix 4. Distributions of outcome measures and covariates in final data

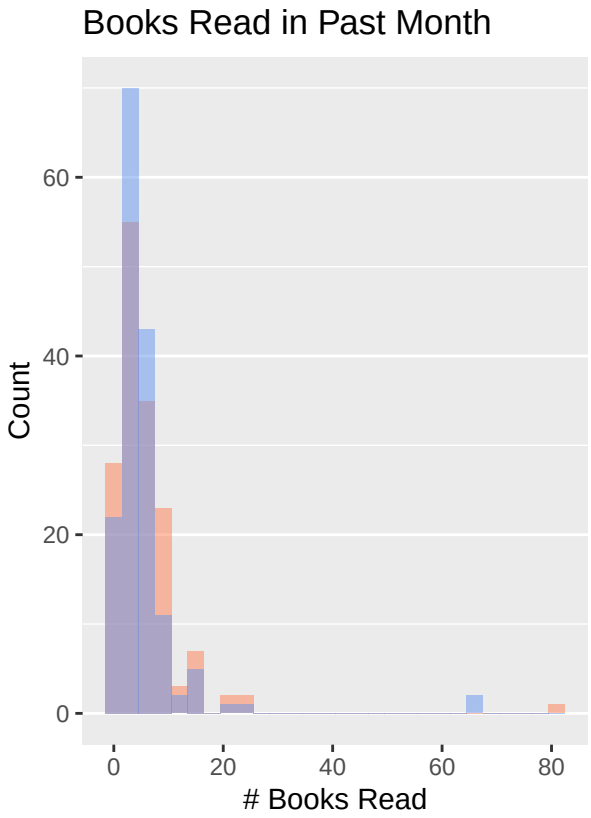
Outcome Measure Distributions

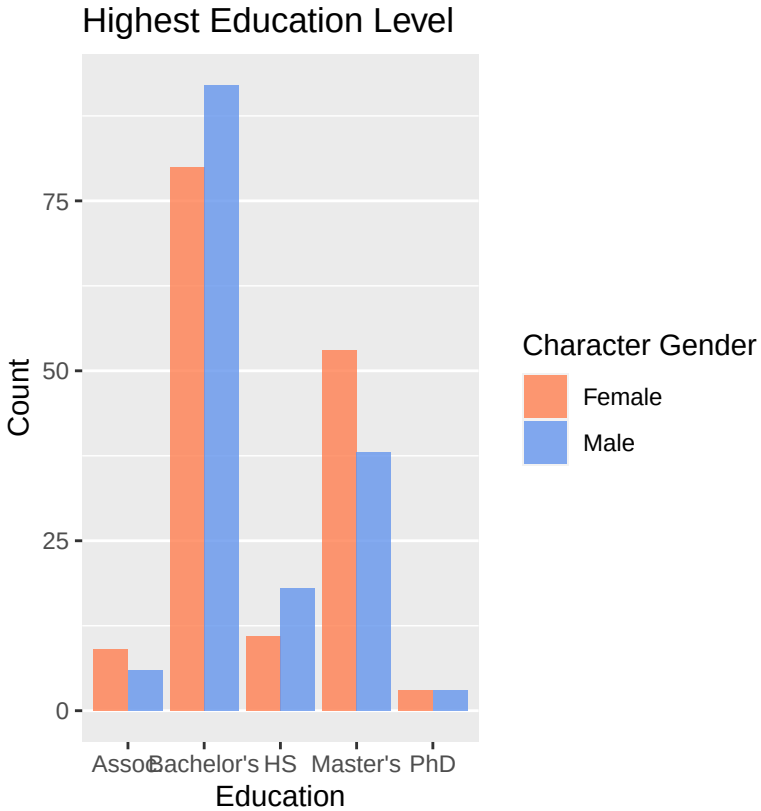




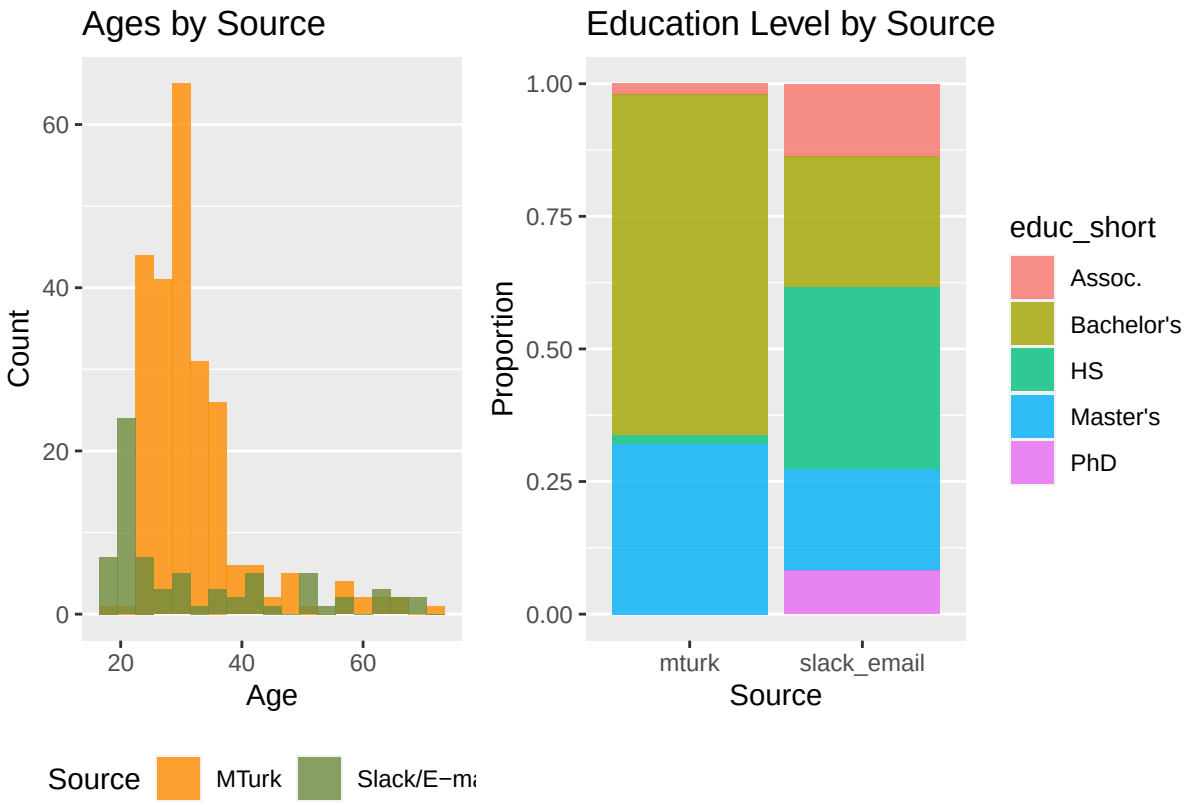
Covariate Distributions


Age Distribution
Count
30
20
10
0
20
40
60
Age


Books Read in Past Month
Count
60
40
20
0
0
20
40
60
80
Books Read


Highest Education Level
Count
75
50
25
0
Assoc
Bachelor's
HS
Master's
PhD
Education
Character Gender
Female
Male

Appendix 5: Exploratory visualizations of trends between covariates and source



Works Cited

Fiske, Susan T. “Stereotype Content: Warmth and Competence Endure.” Current Directions in Psychological Science, vol. 27, no. 2, Apr. 2018, pp. 67–73, doi:10.1177/0963721417738825.

Fiske, Susan T., et al. “A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition.” J. Pers. Soc. Psychol., vol. 82, no. 2, 2002, pp. 878-902.

Flatt, Molly. “Is the Future Female? Fixing Sci-Fi’s Women Problem.” The Guardian, Guardian News and Media, 29 Aug. 2018, www.theguardian.com/books/2018/aug/29/is-the-future-female-fixing-sci-fis-women-problem.

Konnikova, Maria. “Do Readers Judge Female Characters More Harshly Than Male Characters?” The Atlantic, Atlantic Media Company, 7 May. 2018, https://www.theatlantic.com/sexes/archive/2013/05/do-readers-judge-female-charactersmore-harshly-than-male-characters/275599/.

Mayo, Margarita. “To Seem Confident, Women Have to Be Seen as Warm.” Harvard Business Review, 8 Jul. 2019, https://hbr.org/2016/07/to-seem-confident-women-have-to-be-seen-as-warm.

Menadue, Christopher Benjamin, and Susan Jacups. “Who Reads Science Fiction and Fantasy, and How Do They Feel About Science? Preliminary Findings From an Online Survey.” SAGE Open, Apr. 2018, doi:10.1177/2158244018780946.