W241 Final Project

Amar Chatterjee, Kate Lund, Karthik Rameshbabu, Sarah Xie

August 4, 2021

Contents

1 Abstract

As technology continues to introduce new standards for virtual communication and human interaction, remote working and online dating have become the norm. With such a heavy dependency on virtually relaying information, people may have become more skeptical of whom or what they interact with from behind a screen. In this study, we explored the impact of non-verbal communication on interpersonal trust by exposing participants to a case study with questionable veracity. Participants in our control group were presented with the details of the case via only audio on Zoom, while participants in the treatment group received the same audio with accompanying video. A proxy variable for the participant's confidence in the truthfulness of the case was measured in both groups. We found that participants who received the information via both audio and video were not statistically significantly different than participants who received the information via only audio, suggesting that there may not be a link between trustworthiness and non-verbal communication as it was presented in this experiment. We did also find interesting correlations and trends from our data that could drive future research.

Please note: This document uses in-line hyperlinks to point to references. You are free to click on any of the links to access the relevant materials. You are encouraged to be connected to the internet before doing so.

2 Background

Body language has been shown to illustrate one's confidence and commitment. It involves a host of nonverbal cues or signs such as body movements, facial expressions, tone of voice, and gestures in communication. The importance of body language is that it assists us in understanding what a person is saying, specifically in interpreting moods and emotions.

In his 1971 publication, *Silent Messages*, body language expert Albert Mehrabian concluded that body language accounts for 55% of effective communication. By engaging in more effective communication, we expected subjects from the treatment group, when deciding on the truthfulness of a case, to exhibit higher levels of confidence in their decisions.

In a 2002 study by Day et al, a comparison between therapy patients was made to see if there were differences in the treatment effect of face-to-face sessions versus video or audio virtual meetings. While the overall conclusion was that improvement in symptoms was consistent across all treatment mediums, there were several differences that we found noteworthy. First, patients in audio and video sessions were actually *more* engaged than their face-to-face counterparts. On the other hand, the face-to-face participants were found to be less hostile. While measuring the effectiveness of psychotherapy is not directly related to our exploration of trust in an exchange, the fact that there were measurable differences in the participants' behavior begs the question of whether trust or confidence from that exchange would also be influenced.

As discussed above, it is well-documented that communication is made up of more than just words and how they are conveyed. We sought to understand what impact, if any, that the method of communication has on overall trust during an exchange. This trust was measured via a monetary wager of the participant's choosing which indicated their confidence in the truthfulness of what they were told following the exchange.

3 Research Question

Do people have a higher level of confidence when judging the truthfulness of a narrative, when delivered remotely, via an audio and video combination versus audio only?

4 Hypothesis

We hypothesized that the video treatment would cause an increase in confidence when determining the truthfulness of a story. The mechanisms that made up the treatment break down into visual cues such as body language and other non-verbal communication that occurred as the video was presented. We controlled for other potential confounding variables by maintaining consistency among other facets of the experiment. These included: attire/appearance, intonation, names (nondescript Zoom names), eye contact, posture, moderator(s), and more. By pre-recording the content that was shown both visibly and audibly, the study took great care in standardizing these mechanisms to minimize the effect of confounding variables. Lastly, the individual we requested to do the recording was very far removed from our immediate networks, so this helped to avoid any bias that may have come from knowing the speaker.

5 Experiment Design

5.1 Experiment Overview

The experiment was designed to measure the effect of non-verbal communication on trustworthiness by subjecting study participants to a recording of a short narrative via Zoom and comparing results across subjects in the control and treatment groups. This was a between-subjects design. Subjects were randomly

assigned to either treatment (received video recording) or control (received only the audio recording), and then asked to speak aloud as they processed whether they believed the story to be true or false. To further measure the confidence in their decision and their trust in the person who delivered the narrative, study participants were asked to wager between $0 and $5 (in one dollar increments) in addition to providing their guess. The baseline compensation for the study was $5, and the amount of the wager was either incremented to or decremented from the baseline compensation depending on whether the participant guessed correctly. The correct answer was that the narrative is true.

Some snippets of the script are provided below. A link to the full script can be found in the Appendix.

"Hi there! In a moment I am going to read a true story to you. You will then be given instructions on how we'd like you to respond to the information presented."

"Now that you have listened to the narrative, we would like for you to tell us whether you believe it to in fact be true, or false."

"For starters, you have earned a baseline $5 for your participation in the study thus far. Congratulations! You may now choose to wager a portion of that compensation while answering whether you believe the story to be true or false."

5.2 Project Timeline

The initial sign-up form was sent out to potential participants on Jun 16, 2021, with the first study conducted on Jun 21, 2021. The study was run for a period of 4 weeks, with an additional recruitment sprint after two weeks of running the study to boost the number of participants. Analysis and compilation of the study's results were completed in the 3 weeks following the study, July 16, 2021 - August 6, 2021.

5.3 Measurement Tooling

During study sessions there were two moderators on call with the participant. While the Lead Moderator guided the participant through each session, the second Moderator was responsible for silently capturing data points from the session in the background. These included: number of questions the participant asked before and after the narrative recording began, time taken to make a decision, number of filler words, their wager amount, and the guess made. Detailed descriptions for each of these data points can be found in the Appendix. In addition, the research team captured whether the subject smiled or not at a very specific timestamp as part of the narrative they received. All of this data was captured in a shared Google spreadsheet, and payments were made at the conclusion of each day to all of the participants from that day (as applicable).

We assigned Moderators to each subject using the following order of operations: Do any of the Moderators know the subject? If so, they cannot be the Lead Moderator. Which Moderators are available to conduct the study, taking into account personal & work commitments?

The above methodology worked in assigning Moderators to each subject with almost no issues. In the rare event no Moderators were available to conduct the study, the researchers asked the participant to reschedule to a new date.

This hands-on method of measuring ensured a high level of compliance since moderators were directly in charge of administering the treatment, and helped reduce variability while administering.

5.4 Communication Tooling

The medium for delivering the experimental study was the Zoom platform. Two of the researchers served as Moderators for each participant. The participants were asked to turn their video cameras on for the duration of the study, while Moderators kept their cameras off to minimize distractions or biases. Moderators

read aloud instructions to each of the participants from a script to minimize variation, and delivered the intervention of an audio (control) or video (treatment) recording via Zoom's sharing capabilities. The recording was exactly the same irrespective of assignment group – the only difference being the addition of the accompanying video for the treatment group.

5.5 Enrollment & Recruitment Process

To make scheduling easy for participants, and manageable for the research team, we used Calendly as a booking platform. This allowed us to recruit subjects into the study, collect pre-study demographic information, generate calendar invites with meeting times, and send meeting reminders to help minimize attrition. (sample link in Appendix) Each of the researchers reached out to their personal networks (friends, family, colleagues, classmates, etc.) via social media, text, email, and other various forums. The Calendly link was also publicly posted on a variety of Slack channels within the UC Berkeley School of Information domain. Potential participants were incentivized to participate with the lure of earning $5 in compensation with the opportunity to double their earnings.

Interested participants were asked to fill out a simple questionnaire with their contact and demographic information, and choose a day & time to participate in the study via Calendly. No information on what the study would entail was provided, other than the requirement to have access to a video camera while participating.

The researchers set up a shared Google account to capture all participant sign ups and coordinate the calendar. The information from the sign up form was transcribed to a shared spreadsheet to capture and record additional information during the study. Moderators were assigned to each sign up slot based on availability, while making sure to take into account that they were not the Lead Moderator if they knew the subject directly.

5.6 Power Study

We conducted our power study using an application shared by one of the course instructors. We found that in order to achieve a 0.5 treatment effect size with a standard deviation of 1 with 60% power, we would need a sample size of 79. We achieved our goal with 80 participants, and observed a positive treatment effect of about **$0.54** with a robust standard error of **$0.47**. However, this result was not statistically significant as evident in the exploration below, signaling that our experiment was underpowered.

5.7 Randomization

Randomization is a crucial step in conducting a strong experiment, so the team was careful when assigning subjects to the control and treatment groups. First, participants were blocked based on their gender and their ethnicity, which was provided at the time of signing up for the experiment. We believed these two factors to have the strongest potential impact on a subject's confidence level. We chose not to block on any additional covariates given our limited sample size.

After separating out the participants, we ended up with four blocks: non-female and Caucasian, non-female and non-Caucasian, female and Caucasian, and female and non-Caucasian. The gender and ethnicity were split into two groups each (female vs. non-female and Caucasian vs. non-Caucasian) because the speaker who presented the case was both female and Caucasian. These are both clearly discernible characteristics in the video treatment that could have had an impact on the participants' responses. One downside to this blocking technique was that we ended up with some very small sample sizes in some of the groups, as seen below:

Table 1: Random Blocking Assignment

Gender	Ethnicity	# Control	# Treatment
Female	Caucasian	5	6
Female	Other	18	18
Other	Caucasian	3	5
Other	Other	12	13

Within each block, participants were placed in the treatment or control group using a random number generator in R. The total number of participants (`N`) in the block was counted and half of them were randomly assigned to the control group by selecting `N / 2` subject IDs (a number from 1 to 80) from the list of IDs within that specific block. When `N / 2` was an odd number, the extra subject was assigned by flipping a coin.

It is important to note that this process of assigning subjects to control and treatment groups was done in batches due to the nature of our recruitment process. In order to collect data from as many participants as possible, the team left the sign up link open for several weeks while the experiment was being run. The study was conducted for participants as they signed up on a rolling basis. Therefore, participants were blocked by their gender & ethnicity and assigned to control/treatment groups in chunks before they were scheduled to undergo the experiment. We closed the sign up link after we had received over 80 signups, in accordance with our power study.

This methodology led to slightly unbalanced control and treatment assignments by gender and ethnicity, though we do not believe it influenced the integrity of our experimental design. Random assignment was used throughout the process as detailed above.

5.8 Observations & Outcome Measures

Our main outcome measure for this study was the subjects' wager amounts.This was used as a proxy for a subject's confidence in their response to the case presented to them. The wager amount was limited to discrete dollar intervals between `$0` and `$5`. In addition, the team collected a variety of additional data points during the experiment, many of which could also be used as outcome variables albeit not being the focus of our experiment.

In addition to the primary outcome wager described above, we measured six additional outcome variables described in the paragraphs below.

During the session we ask the participants to judge whether the story is true or false with some justification for the answer and their wager amount. While they talked through their logic, the second moderator was timing the length of the response as well as the number of filler words we defined as things like "uh," "um," "er," and the like. The time was measured in seconds and the number of filler words were integers. The thought was that perhaps the more confident people were, the less filler words they would use, or perhaps the less justification they felt they needed to provide.

At one point during the recording, the presenter of the case becomes animated and smiles widely at the camera,and says, "you've earned a baseline $5 for your participation in the study thus far. Congratulations!" We recorded whether the participant smiled at that point, this was measured a boolean, 0 for no the participant did not smile and 1 for did smile. While not directly related to our initial research question, we considered that even if the treatment delivery of the narrative did not affect the confidence in the truthfulness, there may be other social impacts like engagement or personability possibly indicated by a participant smiling or not.

We noted whether the participant ultimately answered correctly. The case we presented was actually true, despite its seemingly incredible circumstances. If the participant reported they thought the story was true we recorded a 1, and if they reported it was false, a 0. While we ruled this out as a proxy for confidence, we

again thought it would be interesting if the treatment group was more able to discern the truthfulness than the control group.

There were a variety of justifications provided by participants for the ultimate assessment of the truthfulness of the story and their wager amount. We looked for inclusion of historical context and other background information to justify their response and classified that as a logic based argument. The implications of this will be discussed in a later section. Participants that relied on a gut feeling, or identified that in the recording they were in fact told at the beginning that it was a true story did not use logic and were recorded as 0.

The final outcome variables we recorded were the number of questions asked pre and post delivery of the recording, this was initially added to provide indication if our instructions and script were sufficiently clear for the participants. While we maintained the measurement throughout the study, we ultimately decided not to use it as an outcome measurement in the analysis phase as very few people had any questions. The measurement was an integer value of the number of questions.

A detailed list of all variables with their descriptions can be found in the Appendix.

6 Results

6.1 Data Completeness

In total, the dataset contained 80 rows and 26 columns, one row for each participant. The data was very complete since each participant was required to fill out the entire survey for collecting personal information when signing up for the study.

6.2 Data Cleaning

The team took great care in setting up the data collection process so that it would remain as error-free as possible. Because the team conducted every single experiment session we were able to quickly flag any problematic observations, therefore the data did not require many cleaning steps post-collection. After exporting the data to a CSV, we conducted the following data manipulations in Python to prep the data for analysis in R.

Cleaning:

- Dropped all columns that revealed any private/personal data.
- Dropped 1 row that was a bad study sample due to poor internet connection.

Transformations:

- Cleared $ signs and converted all monetary value columns to type float.
- Converted non-monetary numeric columns to integers.
- Transformed all date string columns to datetime format.

Creating Derived Data Columns:

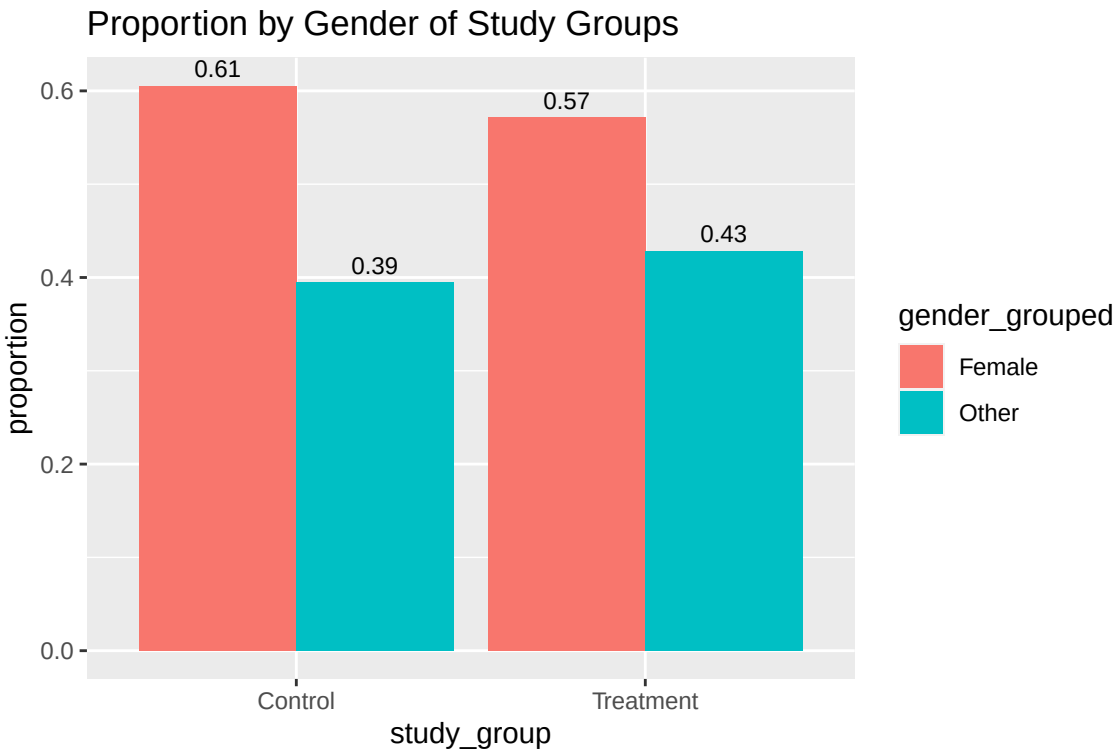
- **isTreatment** (boolean): **1** if study group was treatment, **0** if study group was control
- **isFemale** (boolean): **1** if gender was female, **0** if gender was not female
- **isCaucasian** (boolean): **1** if ethnicity was Caucasian, **0** if ethnicity was not Caucasian
- **is_or_was_married** (boolean): **1** if marital status was married, **0** if marital status was anything other than married
- **uid** (nominal): randomly generated unique identifier to remove sensitive data and avoid any internal bias the team may have when conducting analysis
- **start_hod** (ordinal): hour of day the study was conducted in PDT
- **local_hod** (ordinal): hour of day the study was conducted in participant's local timezone
- **word_per_sec** (numeric): count of participant's filler words (**count_filler_words**) divided by total time taken to answer (**seconds_taken_to_decide**)

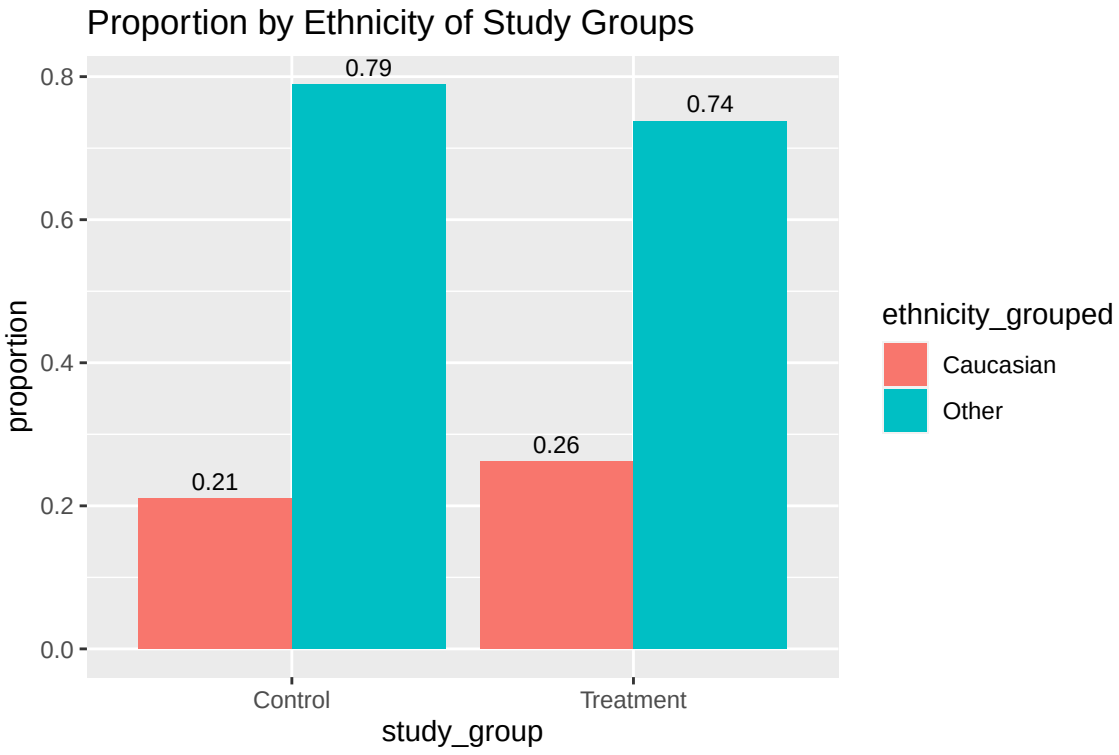
A detailed list of all variables with their descriptions can be found in the Appendix.

6.3 Randomization Check: Covariate Balance

To check covariate balance between the control and treatment groups, we compared the proportions of the groups within each participant characteristic. The characteristics recorded were gender, age range, ethnicity, highest level of education, marital status, employment status, US region, and referral method. Please see the Appendix for the full list of charts.

As mentioned above in the Experimental Design section, we blocked the control and treatment groups by gender and ethnicity, so both the gender and ethnicity (Caucasian vs. non-Caucasian) features were well-balanced between the control and treatment groups.





The remaining covariates, however, showed varying degrees to differentiation between the control and treatment groups. While age range, marital status, and US region were relatively balanced, employment status and referral methods differed noticeably. This could suggest that the treatment and control groups represent

fundamentally different types of subjects and thus we are unable to make an apples-to-apples comparison between them. However, due to the careful random assignment process we undertook and our relatively small sample size, some imbalance is to be expected and likely to have minimal impact because there is no room for selection bias. We also believe these features to be less impactful to our outcome variable, and keep this imbalance in mind when interpreting our results.

7 Regressions

We ran a series of regression analyses on both the main outcome variable, `wager_amount`, as well as the secondary outcome variables, `seconds_taken_to_decide`, `words_per_sec`, `did_smile`, and `logic_based_argument`.

7.1 Primary Outcome Variable

In order to estimate the treatment effect of watching the video versus hearing audio only on the participants' confidence in the truthfulness of the story, we first ran a simple linear regression of our outcome measure (wager amount) on a binary flag indicating whether the participant was in the treatment or the control group. The output of this regression is depicted in Table 2.

Null Hypothesis: The treatment of receiving the video causes no change in wager amount compared to the control of receiving just audio.

Table 2:

	Dependent variable:
	wager_amount
is_treatment	0.541
	(0.466)
Constant	2.316***
	(0.338)
Observations	80
R^2	0.017
Adjusted R^2	0.005
Residual Std. Error	2.055 (df = 78)
F Statistic	1.385 (df = 1; 78)
Note:	*p<0.1; **p<0.05; ***p<0.01

The resulting regression output suggests that receiving the video case increases wager amounts by 0.541 (0.466) as compared with the audio case. This estimate is not statistically significant at the 5% level due to the p-value being greater than 0.05. This suggests that the difference we see in our treatment versus control may be simply due to random chance, and we do not have enough evidence to reject the null hypothesis.

Next, we ran a series of regressions of the wager amount on other covariates such as gender, ethnicity, and marital status. The full analysis is displayed below, in Table 3 and Table 4 (full model).

We find that ethnicity, which here is represented by a binary flag of whether the participant is Caucasian or not, is statistically significant at the 10% level, suggesting that we would encounter an estimate as extreme as 0.966 (0.57) purely due to random chance only 10% of the time. Therefore, while we fail to reject the null hypothesis that the treatment (video case) has no impact on confidence in the truthfulness of a story, we

Table 3: Wager Model Results

	Dependent variable:			
	wager_amount			
	(1)	(2)	(3)	(4)
is_treatment	0.541	0.538	0.485	0.484
	(0.466)	(0.473)	(0.476)	(0.480)
is_female		−0.095	−0.088	−0.111
		(0.480)	(0.476)	(0.479)
is_caucasian			1.041*	1.023*
			(0.557)	(0.564)
is_married				0.269
				(0.476)
Constant	2.316***	2.373***	2.150***	2.040***
	(0.338)	(0.454)	(0.463)	(0.504)
Observations	80	80	80	80
R^2	0.017	0.018	0.065	0.069
Adjusted R^2	0.005	−0.008	0.028	0.019
Residual Std. Error	2.055 (df = 78)	2.068 (df = 77)	2.031 (df = 76)	2.040 (df = 75)
F Statistic	1.385 (df = 1; 78)	0.704 (df = 2; 77)	1.752 (df = 3; 76)	1.388 (df = 4; 75)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: Full Wager Model

	Dependent variable:
	wager_amount
is_treatment	0.372
	(0.496)
is_female	−0.061
	(0.482)
is_caucasian	0.966*
	(0.570)
is_married	0.295
	(0.478)
has_advanced_edu	−0.596
	(0.499)
Constant	2.339***
	(0.575)
Observations	80
R^2	0.089
Adjusted R^2	0.027
Residual Std. Error	2.032 (df = 74)
F Statistic	1.442 (df = 5; 74)
Note:	*p<0.1; **p<0.05; ***p<0.01

have sufficient data to suggest that ethnicity is indicative of wager amount, as only 10% of the experiments conducted in this way would observe this extreme of a difference in ethnicity indicator (Caucasian vs. not Caucasian) arise by chance.

The lack of an observed significant treatment effect could've been due to a number of reasons. First, we observed some behavior from participants that could be classified as noncompliance, namely from never-takers and always-takers. There were participants in the treatment group that did not pay attention to the video they were being shown, as they were looking elsewhere or were too focused on listening to watch the video. These individuals would not have been impacted by anything shown in the video recording. This behavior would likely be consistent within these individuals regardless of whether they were in the control or treatment group and would cause our treatment effect measurement to appear lower than it could be, so we classify them as never-takers. Additionally, there were individuals who felt that $5 (the highest possible wager amount) was not very much money to win or lose, so they wagered $5 regardless of how confident they felt about their response on whether the story was true or false. This is always-taker behavior, because regardless of control or treatment group assignment, these individuals would exhibit the expected treatment behavior and could inflate our treatment effect measurement. Because of these reasons and our inability to isolate the effect of noncompliance, we designate any treatment effect observed from this study as an **intent-to-treat effect**.

While the team did track information on never-taker and always-taker behavior (`logic_based_argument` represents never-taker behavior, and always-taker behavior was written down in our notes), removing these samples from the analysis would've dramatically reduced our statistical power, dropping our total sample size of 80 down to 59. Though the occurrence of never-taker and always-taker behavior was randomly distributed across the control and treatment groups, thus posing a lighter threat to the validity of the study, we would like to err on the side of caution. To the extent that we have accurately characterized the behaviors of these subjects, it is appropriate to label them as non-compliers and to label our reported causal estimate as an intent-to-treat effect.

7.2 Secondary models: Seconds taken to decide

`seconds_taken_to_decide` - Reflects the number of seconds the subject took to decide on their answer (true or false) and wager (if they chose to wager). Distribution depicted in Figure 1.

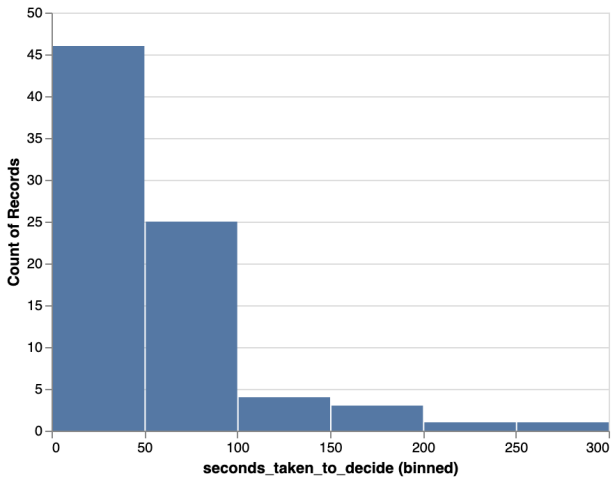


Figure 1: Distribution of seconds_taken_to_decide

Out of all the model variations we ran with this outcome variable, only two models contained any statistically significant estimates.

Table 5: Seconds Taken to Decide Model

	Dependent variable:
	seconds_taken_to_decide
is_treatment	18.641*
	(10.901)
is_female	−8.630
	(11.454)
is_caucasian	−5.015
	(11.954)
factor(start_hod)evening	−6.513
	(17.493)
factor(start_hod)morning	−32.671**
	(13.060)
Constant	67.609***
	(19.182)
Observations	80
R^2	0.150
Adjusted R^2	0.093
Residual Std. Error	44.574 (df = 74)
F Statistic	2.622** (df = 5; 74)
Note:	*p<0.1; **p<0.05; ***p<0.01

In `seconds_model` depicted in Table 5 we regress `seconds_taken_to_decide` on `is_treatment`, `is_female`, `is_caucasian`, and `factor(start_hod)`, which is the time bucket of the day that the participant began the experiment. "Morning" bucket is classified as anything before 10am, "afternoon" is between 10am and 4pm, and "evening" is anything scheduled after 4pm. This model estimates a statistically significant coefficient for the variables `is_treatment` at the 10% level and `factor(start_hod)morning` at the 5% level.

We can see that given the subject was in the treatment, they spent 18.641 seconds more time to decide on their outcome. The standard deviation is somewhat large at 10.901 seconds. We start to ponder about the reasoning behind this, that by seeing and hearing the story the subject tends to ponder and reason more about their answer. When the subject simply hears the story, they may be less invested in the exercise, and may therefore answer more quickly overall.

`start_hod` is originally an ordinal covariate and it is the hour of day (HOD) in Pacific Daylight Time, whose distribution is depicted in Figure 2. It was converted into a categorical feature as described above for this regression. While we did not intend to have the HOD affect our outcome, we realized after-the-fact that subjects may be experiencing different external pressures while participating. Some could be trying to squeeze their timeslot in the middle of their workday, while others may participate during their downtime. The various times of day may inspire different types of moods or behaviors, which in turn may affect the amount of time and effort they are willing to spend. We found that the majority of our participants selected time slots in the morning or late evening, but according to our regression analysis, participants who selected slots in the morning spent a statistically significant lower amount of time, -32.671 (13.06), on the study. While we remain unsure of what the driving force for this behavior is, we acknowledge the impact that the time of day may have on wagering behavior. As the graph below shows, the distribution of participants across HOD is similar across both treatment and control groups due to random assignment, so we do not anticipate any degradation in the integrity of our experiment.

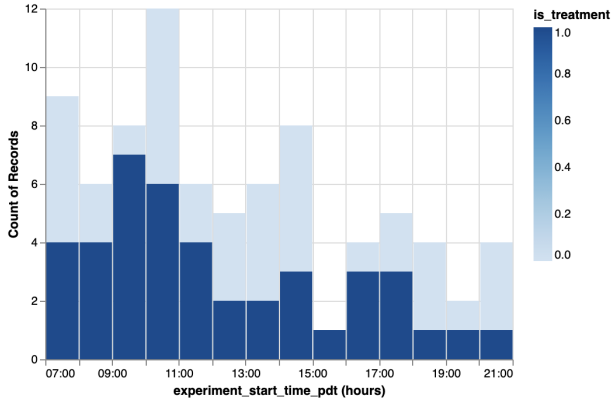


Figure 2: Distribution of start_hod

7.3 Secondary models: Filler-words per second

`words_per_sec` - Reflects the average number of filler words per second that the participant said. This is calculated by taking their total number of filler words and dividing by the total seconds taken to decide. Distribution depicted in Figure 3.

7.3.1 `words_per_sec` EDA:

For the `words_per_sec` models we see statistically significant differences between the control and treatment groups. To reiterate, the outcome measure looks at the average filler words used per second by each subject. In our simple model in Table 6 where we regress `words_per_sec` on `is_treatment`, we see a 0.043 increase

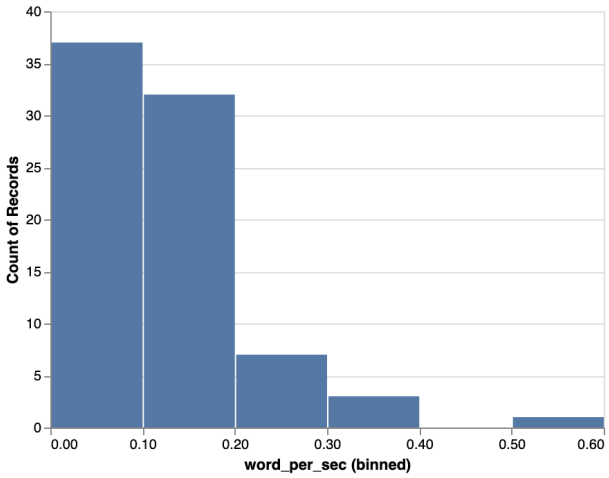


Figure 3: Distribution of words_per_sec

Table 6: Filler Word Model 1

	Dependent variable:
	word_per_sec
is_treatment	0.043**
	(0.019)
Constant	0.101***
	(0.011)
Observations	80
R^2	0.063
Adjusted R^2	0.051
Residual Std. Error	0.085 (df = 78)
F Statistic	5.252** (df = 1; 78)
Note:	*p<0.1; **p<0.05; ***p<0.01

in `words_per_sec` in the treatment compared to the control. We see a strong t-value, paired with a low standard error 0.019 this outcome indicates high confidence in the result. The outcome we see here aligns with the interpretation of our results from the `seconds_model`. When in treatment, visually receiving the narrative may cause subjects to feel more accountable to answer genuinely instead of rushing through the exercise. The longer someone speaks before making a decision, the more filler words they tend to use. In fact we see that `words_per_sec` is positively correlated with `seconds_taken_to_decide`, (corr_coeff = 0.78).

When using gender as a covariate we learned that it did not seem to play a role in the outcome measure of `words_per_sec`.

Table 7: Filler Word Model 2

	Dependent variable:
	word_per_sec
is_treatment	0.041^{**}
	(0.018)
is_female	0.002
	(0.019)
is_caucasian	0.043
	(0.030)
Constant	0.091^{***}
	(0.019)
Observations	80
R^2	0.108
Adjusted R^2	0.073
Residual Std. Error	0.084 (df = 76)
F Statistic	3.061^{**} (df = 3; 76)
Note:	*p<0.1; **p<0.05; ***p<0.01

Lastly, the narrator we chose for the recording of our case was a Caucasian female. We collected ethnicity as a data point for every study subject which resulted in about 5 racial groups. However, we felt that since our narrator was Caucasian, deriving a pre-treatment variable for ethnicity that was binary (`is_caucasian`) rather than the initial split by each group would give us better explainability. When using is_caucasian as a covariate in addition to is_female and is_treatment, we see a statistically significant treatment effect of 0.043 at the 10% level with a high t-value of 1.897. This regression is recorded in Table 7.

7.4 Secondary models: Participant Smiling

One of our outcome measurements was a boolean indicator for whether the participant smiled or not when the presenter in the recording enthusiastically tells them they have earned $5 for their participation in the study thus far, congratulations! While we may not have found any statistical link between treatment group assignment and wager amount, there are other mechanisms about interfacing remotely that could be interesting future study directions. By regressing `did_smile_congrats` on `is_treatment`, `is_caucasian`, and `is_female`, we found that participants who received treatment (saw the video) in fact smiled more than the control group (audio only), as shown in Table 8.

This indicates that we may reject the null hypothesis at the 10% level that there was no difference in the prevalence of smiling between treatment and control groups. The suggestion is that the intent-to-treat effect

Table 8: Did Smile Model

	Dependent variable:
	did_smile_congrats
is_treatment	0.208*
	(0.113)
is_caucasian	0.050
	(0.134)
is_female	0.061
	(0.114)
Constant	0.269**
	(0.104)
Observations	80
R^2	0.050
Adjusted R^2	0.012
Residual Std. Error	0.494 (df = 76)
F Statistic	1.322 (df = 3; 76)
Note:	*p<0.1; **p<0.05; ***p<0.01

of having video in addition to audio increases an individual's propensity to smile by 0.208 (0.113). More research is required to see if this result can be replicated in another experiment.

8 Additional Thoughts

We collected other data points throughout the experimental sessions that reveal other interesting notes about human behavior in this setting. One such data point was whether or not the participant used a logic-based argument when speaking about whether they believe the narrative to be true or false. We define a logic-based argument as using historical facts or other contextual information to determine the validity of the narrative as opposed to an uninformed guess based purely on intuition (which should be more influenced by non-verbal communication). Another data point is related to participants smiling at a specific time stamp. The analyses for these additional points can be found in the Appendix.

While we did not find a statistically significant impact of non-verbal communication (video recording) on subject confidence, we were able to observe some interesting behavior from study participants as they underwent the experiment. We also recognize that a number of factors could influence our understanding of the observed treatment effect, such as the participants' understanding of the story itself. The term "black gold" was used in the story, and many participants were unaware that this is just another way to refer to oil, the natural resource. This caused many subjects to focus on that small aspect of the narrative and disengage from the rest of the story. We also had some non-native English speakers among the participants who admitted to having a hard time following the events of the story, and who may not have responded as they would have if they had better understood the narrative. Both these phenomena occurred randomly in the control and treatment groups, so while they should not have a significant impact on the treatment effect, we feel it appropriate to call attention to them.

9 Conclusion

The results of this analysis indicate that more experimentation and analysis needs to be done to fully understand the relationship between non-verbal communication and confidence in truthfulness or trustworthiness. While we failed to find a causal relationship between non-verbal communication and confidence in this experiment, we learned other interesting facts that could be used as a basis for further research. Because we were underpowered for this study, it may be worth repeating the experiment as it is described in this document with a larger sample size to achieve higher power, and determine whether the results here can be replicated.

9.1 Limitations and Future Enhancements

This study began with a few known limitations. Because this was designed and executed within a 14-week school semester, there was realistically only 4-6 weeks to work on this experiment and conduct the study. The options of study within that time frame were limited, and defined our experiment execution and data collection timeline. We were also constrained by our $500 budget which we chose to use as an incentive for participation *as well as* a proxy for confidence, dramatically reducing the amount of funds we could spread across participants. If we had access to more time and financial resources, we would have chosen to perform a pilot study to help work out challenges with the execution of the study before fully rolling it out. We would've also preferred to give our participants more money to work with for their wager, so that we could better simulate the decision-making that surfaces in risky situations and better measure confidence.

As we progressed through the experiment sessions, the team realized a few things about the experimental design and execution that may have elicited unexpected responses from participants – issues that might have been identified and addressed had we run a pilot study. The first was the case we selected to present to participants. We had focused on emphasizing the non-verbal cues in the video case when designing the experiment, and neglected to consider the consequences of the case story itself. As we began to collect participant responses, we realized that many of them focused on the facts presented in the story to form their conclusion, so much so that some participants had their heads down taking notes the entire time, which likely negated much of the impact that watching the video could have. Compounding this phenomenon was the fact that the story we selected turned out to be quite long (about 2 minutes), which caused some participants to lose interest (a study by Microsoft and another by the Technical University of Denmark suggest that the average adult attention span is anywhere between 8-15 seconds and it continues to narrow). Thirdly, our participants were recruited from our personal and professional networks, so there were higher rates of logic-based, educated thinkers than would likely be found in the general population. All of these issues could have impacted how the treatment was received and biased results. If performed again, we would choose a shorter case with less facts mentioned throughout and recruit participants from a more varied pool.

Along the same vein, we believe that the video case could have been more dramatic in order to capture the viewers' attention (and trust). The speaker stayed relatively still in our video, but it may have been helpful to encourage her to make more head or hand motions as she talked, and to smile more, especially when emphasizing key points in the video. This may have increased the treatment effect and allowed us to observe the effect more closely despite our small sample size.

Another component of the experiment that we would likely change (or remove altogether) if this experiment were to be repeated is the request of participants to process their thoughts aloud and explain the reasoning behind their guess (true/false) and wager amount. As we held more and more experiment sessions, it became apparent that this process (of describing their thought process) was causing participants to feel the need to justify their decision and rely purely on facts or logic instead of going with their gut. Some participants even changed their guess or wager amounts midway through the session because they essentially talked themselves into believing the story was false (or true). This may have had some unintended consequences on our treatment effect, but this impact was constant across treatment and control groups so we have no reason to suspect a threat to the internal validity of the study.

10 Appendix

10.1 Modeling Variable Glossary

- `wager_amount`: discrete dollar amount that a participant bet on their assessment of the truthfulness of the story, range of 0-5.
- `did_smile_congrats`: boolean metric for whether the participant smiled when the presenter in the recording enthusiastically says they have earned a baseline $5 for their participation in the study thus far, congratulations!
- `seconds_taken_to_decide`: time in seconds from when Moderator 1 tells the participant to go ahead, until the participant stops talking.
- `count_filler_words`: integer amount of filler words used during the participant's rationale for their assessment and wager amount. This includes words like the following: um, uh, er, like, ya. There was some variation between moderators on what exactly counted as a filler word or not.
- `num_questions_pre`: integer amount of questions asked before the recording begins.
- `num_questions_post`: integer amount of questions asked after the recording ends.
- `logic_based_argument`: boolean metric for whether the participant used factual knowledge or historical background in determining the truthfulness of the case.
- `is_treatment`: boolean metric for whether the participant received the treatment or control.
- `is_female`: boolean metric for whether the participant was female or not.
- `is_caucasian`: boolean metric for whether the participant was caucasian or not.
- `words_per_sec`: continuous variable for the count_filler_words normalized by seconds_taken_to_decide.
- `is_married`: boolean metric for whether the participant is currently married or in a domestic partnership.
- `has_advanced_edu`: boolean metric for whether the participant has at least a Master's degree or higher

10.2 Complementary Materials

The following hyperlinks contain additional information and materials used to conduct our experiment.

- **Live Experiment Script**: Document referred to by moderators during each study session. Contains the *Pre-Flight* Checklist.
- **Calendly Booking Page**: Link to the booking system used to schedule participants.
- **Story Recording**: The recording audio was used for the control group, both video and audio was used for the treatment group.
- **Story Recording Script**: Script that our reader used to film the recording.
- **Email Templates**: Templated outreach messages shared by the research team to their respective networks to recruit participants.
- **FAQ Standardized Responses**: Answers to anticipated questions from participants to maintain consistency across moderators.
- **Pilot Data**: Initial data points and analysis done as study sessions were being conducted.
- **Slide Deck**: Slides used for the final in-class presentation of our research study.
- **Project Timeline**: Gantt chart that we used to track our project progress.
- **Data Glossary**: Descriptions of variables captured and presented in the study.

10.3 Other Models

10.3.1 Did the participant get the answer correct?

In this section we are investigating whether there was a difference between treatment and control groups in judging whether the case presented was true or false. We regressed answer_correct on is_treatment, is_caucasian, and is_female and did not find a statistically significant coefficient, results shown below.

Table 9: Answer Correct Model

	Dependent variable:
	answer_correct
is_treatment	0.132
	(0.116)
is_caucasian	−0.026
	(0.135)
is_female	0.112
	(0.117)
Constant	0.359***
	(0.113)
Observations	80
R^2	0.029
Adjusted R^2	−0.010
Residual Std. Error	0.505 (df = 76)
F Statistic	0.748 (df = 3; 76)
Note:	*p<0.1; **p<0.05; ***p<0.01

In discussing possible proxy variables for the participants' confidence in their assessment of the case presented being true or false, we considered using the answer_correct as the proxy variable. However, given true or false is a boolean outcome, we did not think it would accurately measure the confidence in a participant's response.

10.3.2 Did the participant use a logic based argument?

Here we looked at the effect of the assignment group on using a logic based argument.

However when we ran a regression of this binary variable on the treatment indicator variable, we found an estimate of **0.099 (0.0998)**, which is not statistically significant at the 95% confidence level.

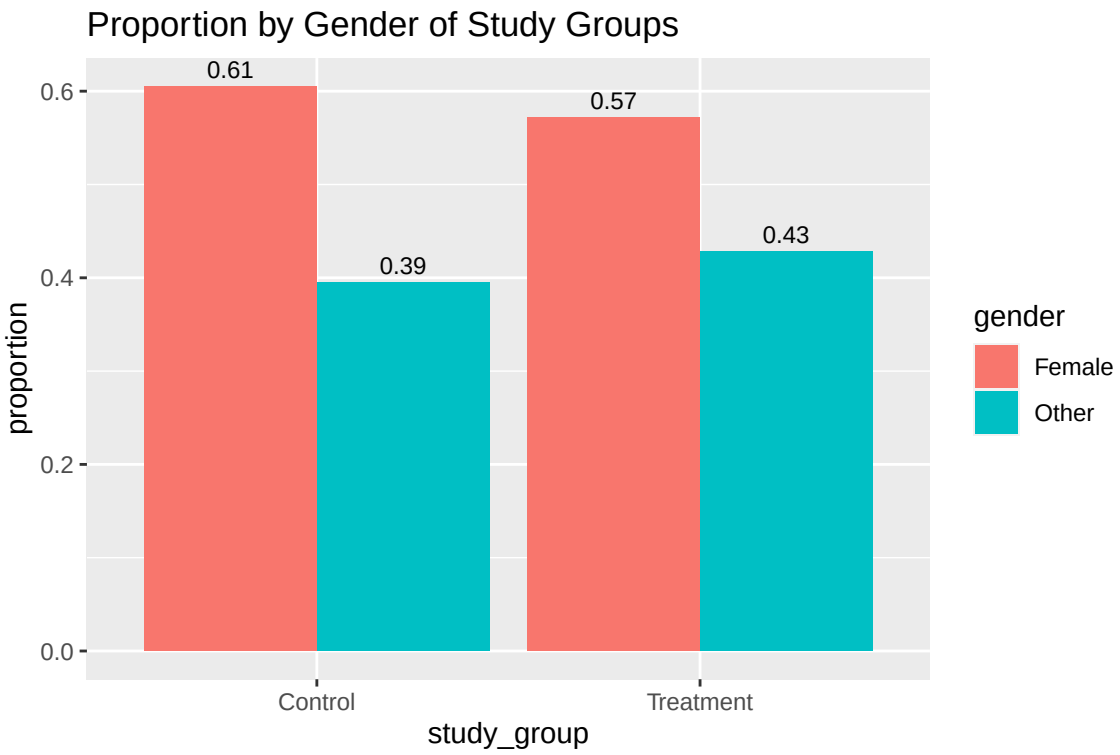
Table 10: Logic-based Argument Model

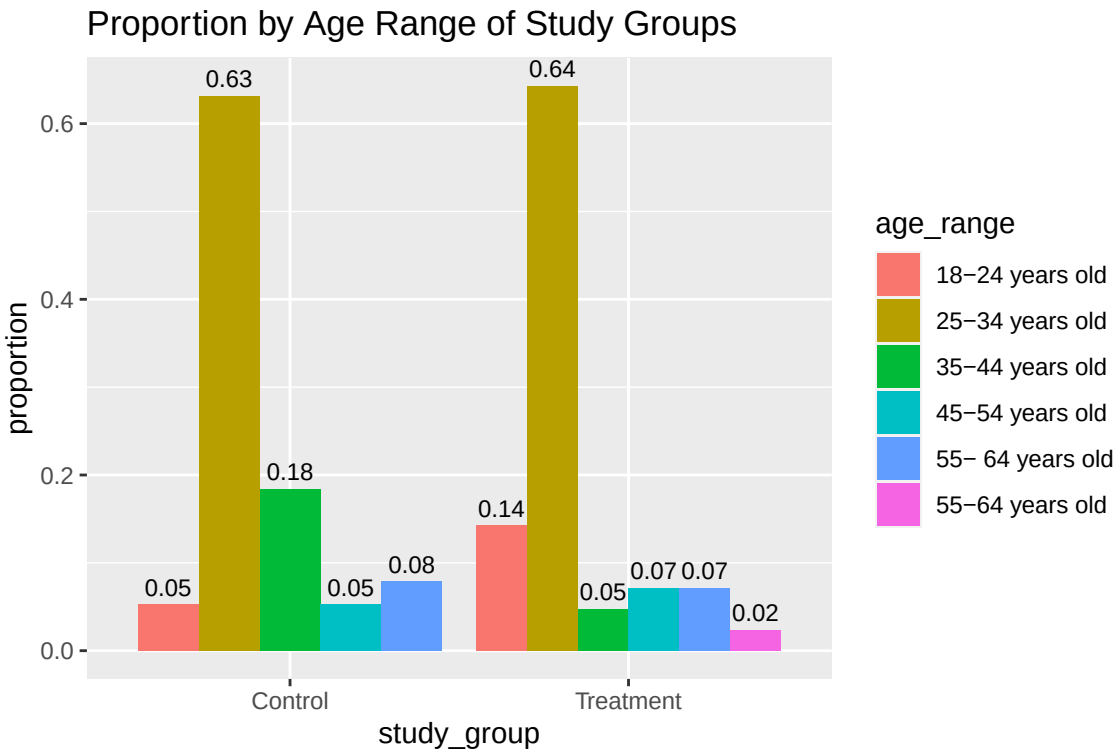
	Dependent variable:
	logic_based_argument
is_treatment	0.099
	(0.100)
Constant	0.211***
	(0.068)
Observations	80
R^2	0.013
Adjusted R^2	−0.00003
Residual Std. Error	0.443 (df = 78)
F Statistic	0.997 (df = 1; 78)
Note:	*p<0.1; **p<0.05; ***p<0.01

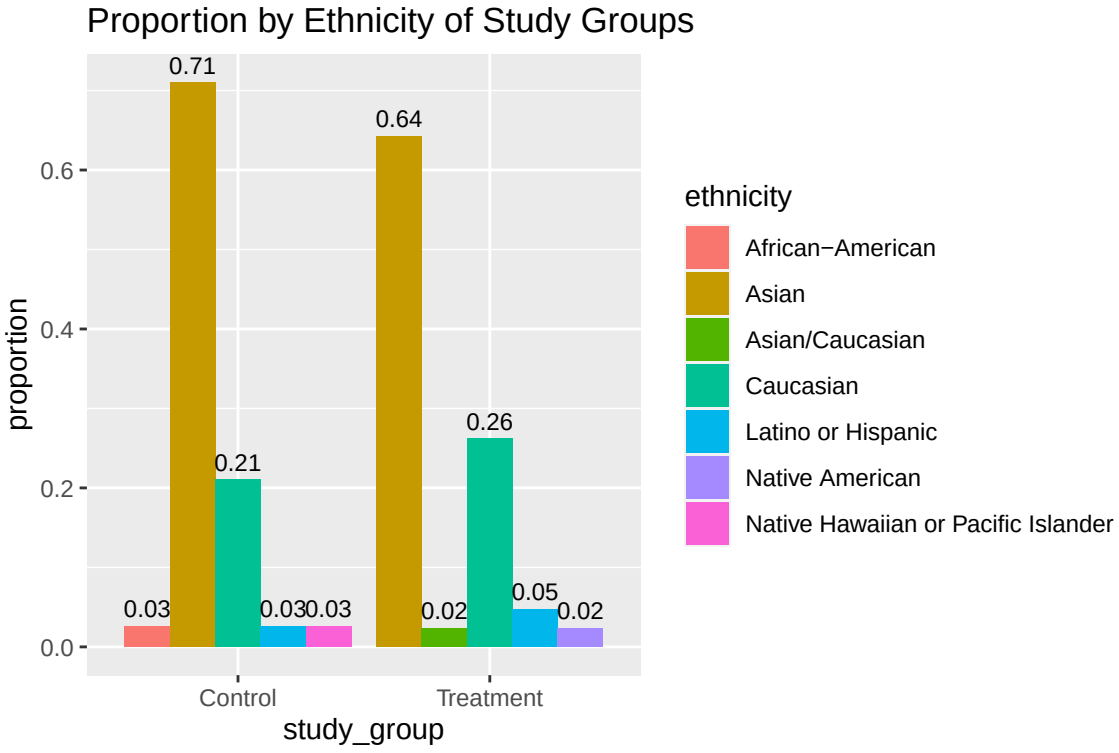
10.4 Visualizations

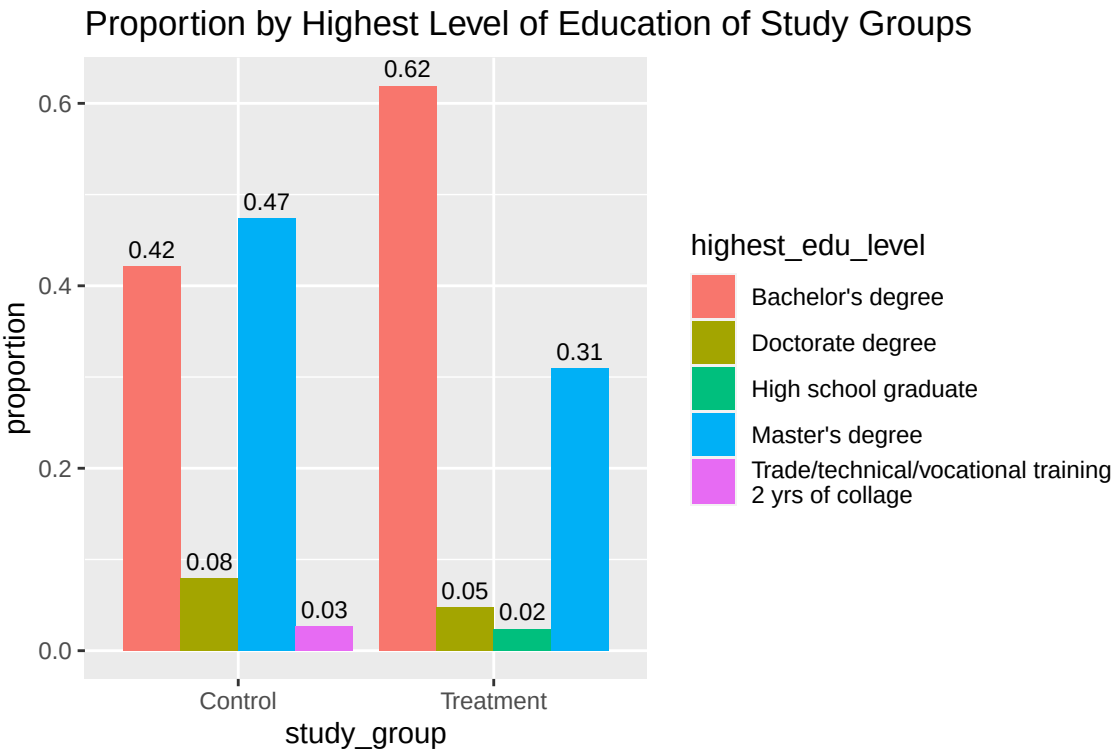
Below are some visual representations of our covariate balance checks, along with some additional charts.

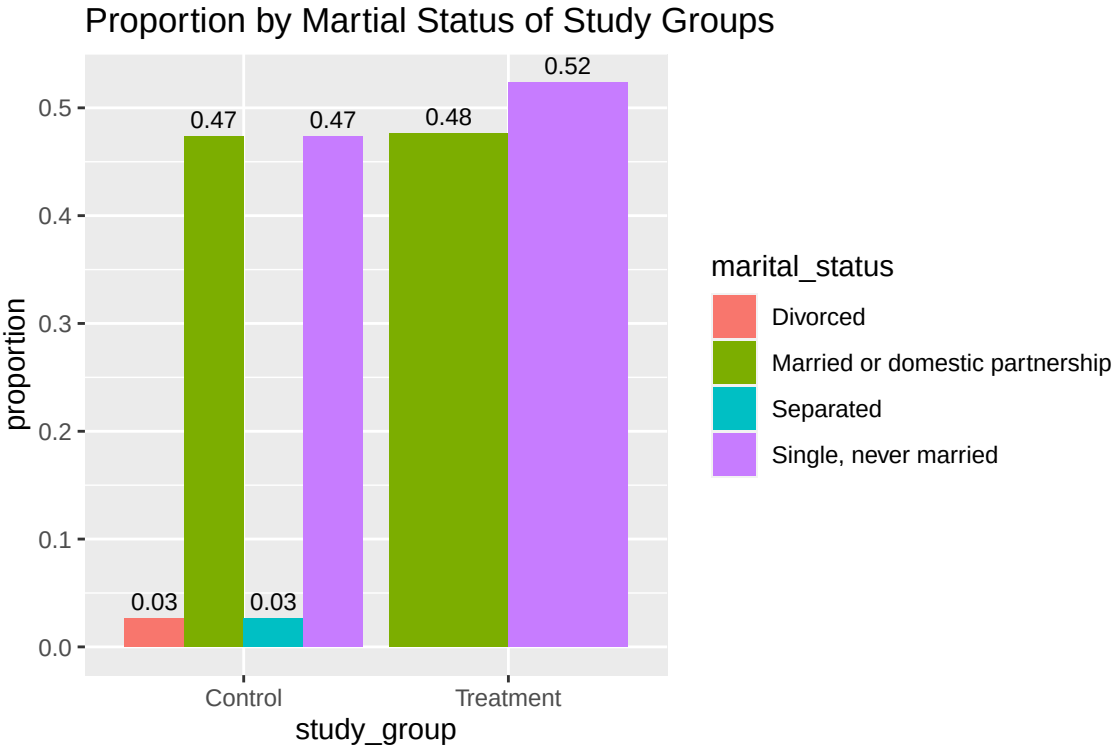
Covariate Balance Checks

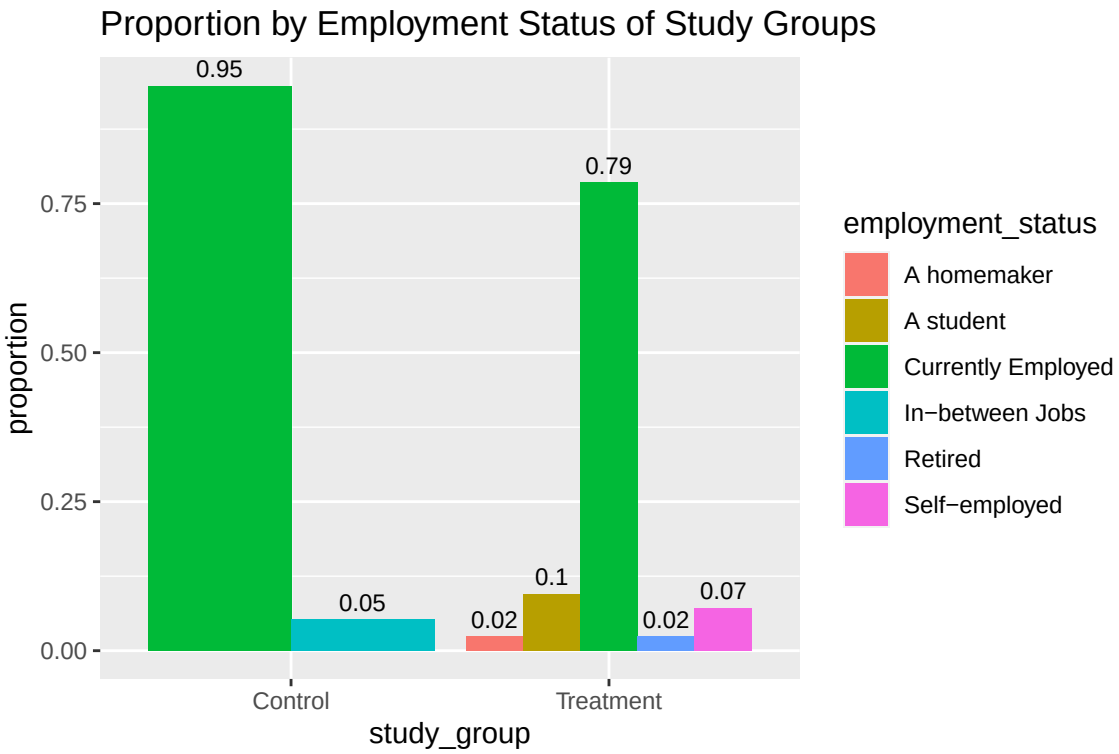


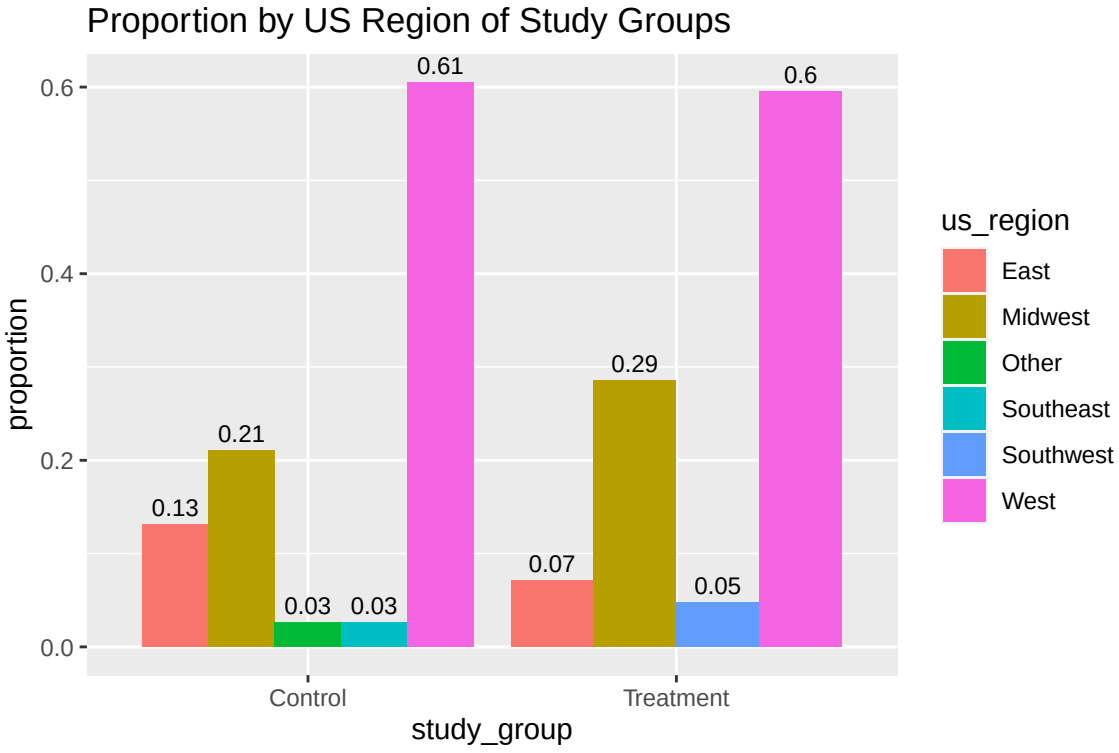

Proportion by Age Range of Study Groups
proportion
0.0
0.2
0.4
0.6
0.05
0.63
0.18
0.05
0.08
0.14
0.64
0.05
0.07
0.07
0.02
Control
Treatment
study_group
age_range
18−24 years old
25−34 years old
35−44 years old
45−54 years old
55− 64 years old
55−64 years old

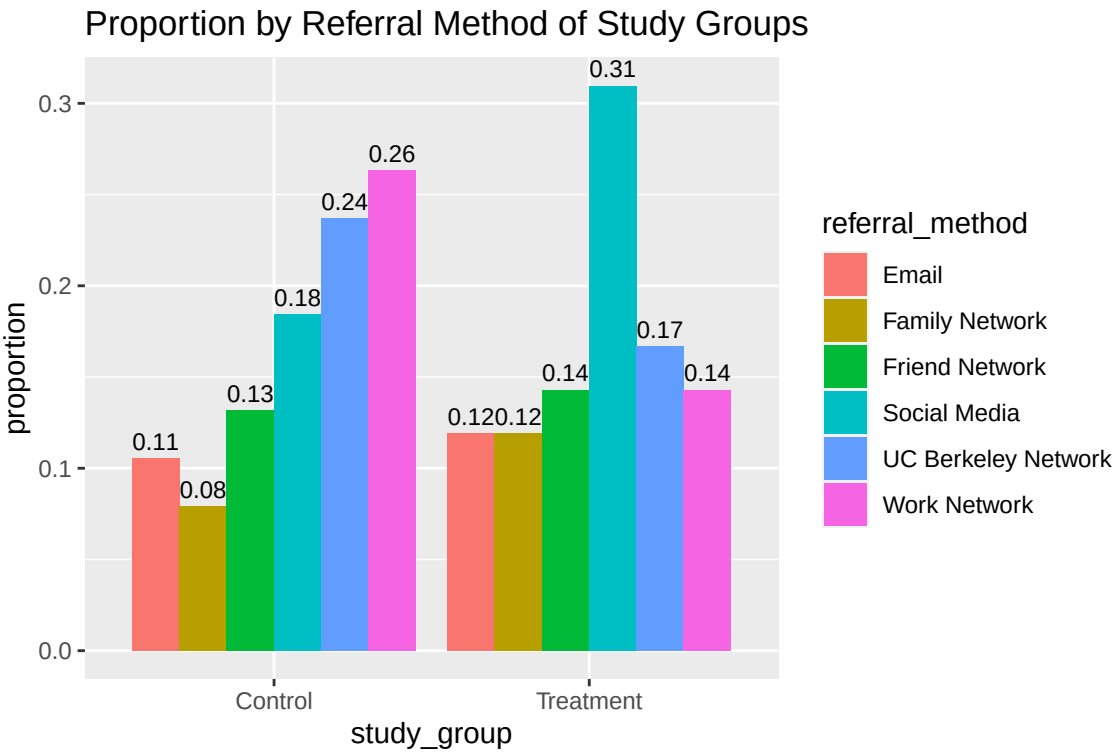

Proportion by Ethnicity of Study Groups
proportion
0.0
0.2
0.4
0.6
0.03
0.71
0.21
0.03
0.03
0.64
0.02
0.26
0.05
0.02
Control
Treatment
study_group
ethnicity
African−American
Asian
Asian/Caucasian
Caucasian
Latino or Hispanic
Native American
Native Hawaiian or Pacific Islander


Proportion by Highest Level of Education of Study Groups
proportion
0.0
0.2
0.4
0.6
0.42
0.08
0.47
0.03
0.62
0.05
0.02
0.31
Control
Treatment
study_group
highest_edu_level
Bachelor's degree
Doctorate degree
High school graduate
Master's degree
Trade/technical/vocational training
2 yrs of collage


Proportion by Martial Status of Study Groups
proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.03
0.47
0.03
0.47
0.48
0.52
Control
Treatment
study_group
marital_status
Divorced
Married or domestic partnership
Separated
Single, never married


Proportion by Employment Status of Study Groups
0.95
0.79
0.05
0.02
0.1
0.02
0.07
0.75
0.50
0.25
0.00
proportion
Control
Treatment
study_group
employment_status
A homemaker
A student
Currently Employed
In−between Jobs
Retired
Self−employed


Proportion by US Region of Study Groups
0.61
0.6
0.13
0.21
0.03
0.03
0.07
0.29
0.05
0.6
0.4
0.2
0.0
proportion
Control
Treatment
study_group
us_region
East
Midwest
Other
Southeast
Southwest
West


Proportion by Referral Method of Study Groups
proportion
0.0
0.1
0.2
0.3
0.11
0.08
0.13
0.18
0.24
0.26
0.12
0.12
0.14
0.31
0.17
0.14
Control
Treatment
study_group
referral_method
Email
Family Network
Friend Network
Social Media
UC Berkeley Network
Work Network

Correlation Plot: `count_filler_words` vs. `seconds_taken_to_decide`

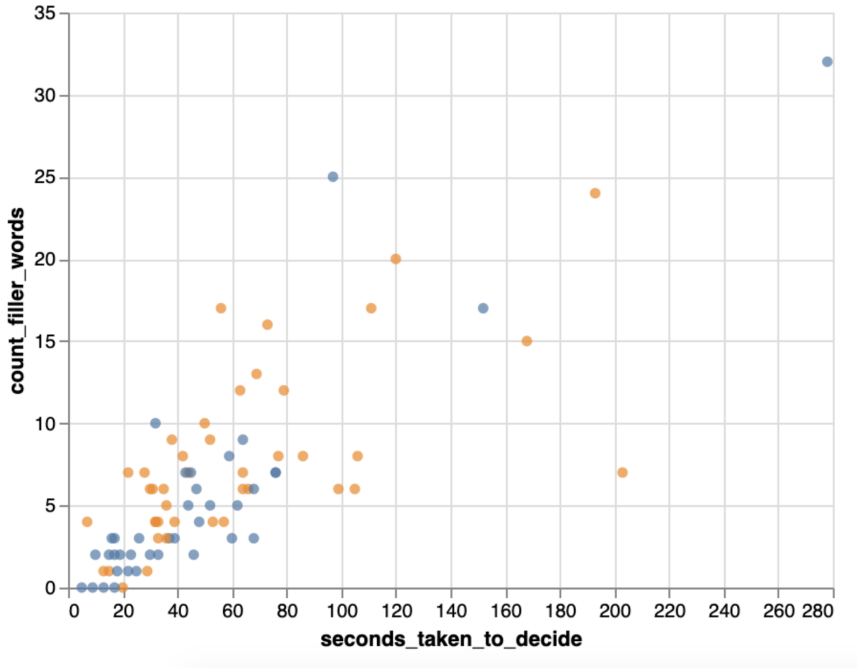
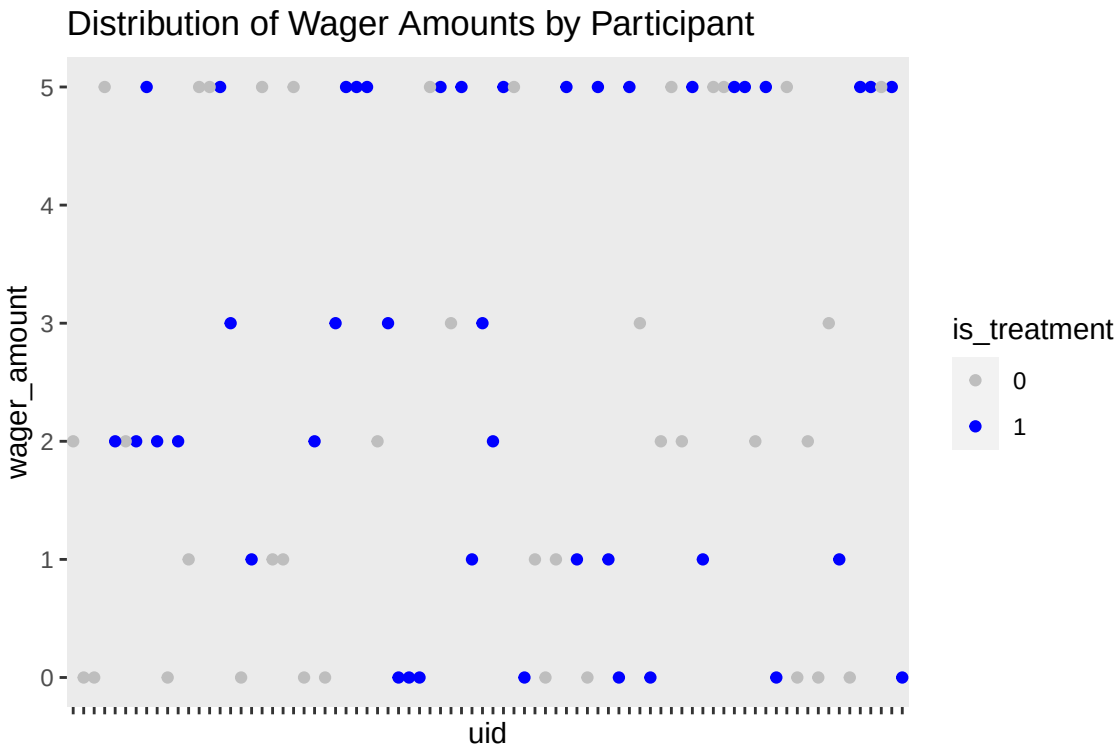


Figure 4: Correlation between filler words and seconds taken to decide

Other Plot:


Distribution of Wager Amounts by Participant
wager_amount
uid
is_treatment
0
1
5
4
3
2
1
0