# Project Analysis

## Introduction

We conducted an experiment to test whether a 6-minutes course can effectively teach participants to recognize manipulative tactics in social media posts. Our experiment consisted of a demographics survey, a pre- and post tests, and the tactics course (or sleep course for the control group). Details of our experiment can be seen in our pre-analysis plan.

### Research Questions

Our experiment focuses on four research questions:

1. Can a 6-minute course focused on manipulative tactics, in the vein of First Draft’s existing SMS courses, help users better identify manipulative content?
2. Does the course reduce the sharing of manipulative content online and offline?
3. Does the course make participants better at identifying each individual tactic covered in the course?
4. Are there heterogeneous treatment effects (HTE) where our course works better for certain subgroups than others? In particular, are there any differences based on users’ misinformation susceptibility at baseline (as measured during the pre-test), political ideology, and income level?

### Hypotheses

These research questions are supplemented with a list of 10 hypotheses.

1. H1: Participants will be more capable of rating misinformation correctly as manipulative after taking the course.
   1. SH1: Participants will not identify true content as more manipulative after taking the course.
2. H2: Participants will be more capable of identifying misleading graphs after taking the course.
3. H3: Participants will be more capable of identifying anecdotes after taking the course.
4. H4: Participants will be more capable of identifying false comparisons after taking the course.
5. H5: Participants will be less likely to share misinformation online after taking the course.
6. H6: Participants will be less likely to share misinformation offline after taking the course.
7. H7: Participants with different levels of susceptibility to misinformation at baseline will react differently to the treatment in terms of their overall ability to identify manipulative content.
8. H8: Participants with different political ideologies will react differently to the treatment in terms of their overall ability to identify manipulative content.
9. H9: Participants with different levels of income will react differently to the treatment in terms of their overall ability to identify manipulative content.

## Summary of Insights

1. Our 6-minutes tactics focused course helped participants label manipulative posts as 12% more manipulative.
2. At the same time, we see increased skepticism for posts in general. Participants also labeled non-manipulative posts as 15.4% more manipulative. In particular, participants were more skeptical of non-manipulative graphs.
3. Participants had difficulty correctly identifying anecdote and false comparison tactics, but our data suggest that participants might have improved x% on being able to identify misleading graphs correctly (our estimate is significant only at the Y significance level).
4. Our tactics course did not seem to change online and offline sharing behavior. We did not explicitly teach about sharing behaviors, so this lack of change is not surprising. We did hope that improved identification of misinformation could lead to reduced sharing behavior. (Note: it is possible that motivation for sharing changed, however, we focus on overall sharing behavior regardless of motivation because all sharing contributes to increasing the reach and popularity of misinformation on social media.)
5. We did not find any HTE effects.

## Data Processing

We will first process the data so that we have all the necessary variables to investigate our hypotheses. We also fix up some column naming typos and values to make them more tractable for analysis.

[Code for preprocessing]

### Attention Check

We choose to filter for respondents who passed both attention checks. This gives us 641 participants, which is 64.4% of our original pool of respondents. See Attention Check Robustness for more details on what happens when we adjust our attention check criteria.

[Code]

## Main Results

### Treatment Effect on Identifying Manipulation

#### H1 - Identifying Manipulativeness

H1 is that: participants will be more capable of rating misinformation correctly as manipulative after taking the course.

Note that, by construction, we test this hypothesis by only looking at test questions that actually contain manipulative tactics. SH1 will examine the test questions that were factual.

[Code for Graph]

The table below summarizes the output of the standard t.test function in R:

\* `estimate` represents the delta in the group means (treated vs. control)

\* `estimate1` is the treated group’s mean delta

\* `estimate2` is the control group’s mean delta

\* `statistic` is the test statistic

\* `p.value` is the unadjusted p-value

\* `p.value\_ajusted` is the p-value with 10 adjustments using the BH correction

\* `conf.low` and `conf.high` are the bounds of a 95% confidence interval

These results indicate that those that take our tactics course rate manipulative posts 0.63 points more manipulative on our 6 point scale. This is 12.6% ($0.63/5$) of our 6 point scale and is significant before and after multiple testing adjustments. This finding is also plotted in the figure above. This suggests that participants became more capable of identifying misinformation after taking the course.

[Code for misinfo t test]

For reference, if we also penalize participants for identifying factual questions as manipulative, then this result is dampened (as illustrated below). This is discussed further with SH1 below.

[Code for complete t test]

#### SH1 - Misidentifying True

SH1 is that: participants will not identify true content as more manipulative after taking the course.

[Graph]

Those that take our tactics course also rate true posts 0.77 points more manipulative on our 6 point scale (15.4% ($0.77/5$)). In other words, our treatment also caused people to be more suspicious of factual information, and so in all information

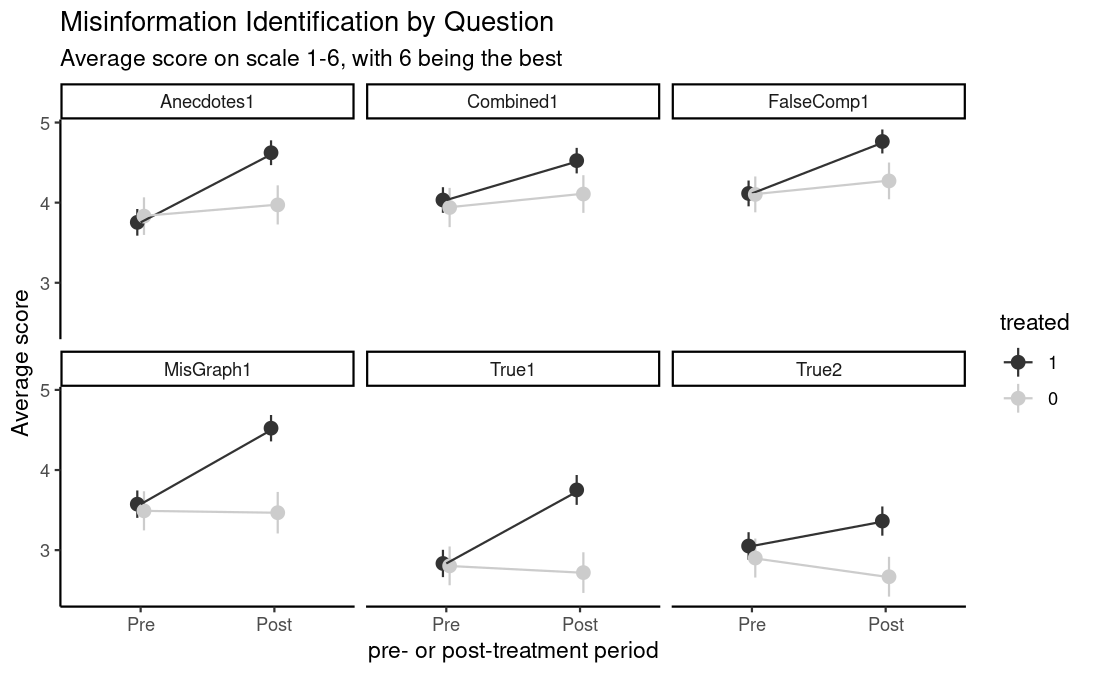
This effect may be a result of priming and might not last as long as the knowledge they gained about identifying misinformation. This result may also be the result of the experimenter demand effect. Thus, we cannot tell apart whether either of these effects is driving the results, or if it is because participants became generally skeptical of all information. In any case, this is something First Draft should be aware of. This finding is plotted in the figure above alongside the result for test questions with misinformation. As illustrated, treated participants became more likely to identify true and false statements as misinformation in the post-test.

[T Test]

#### Breakdown by Question

To get a better understanding of treatment effect on different types of questions, we look at the change in manipulativeness for each question category.

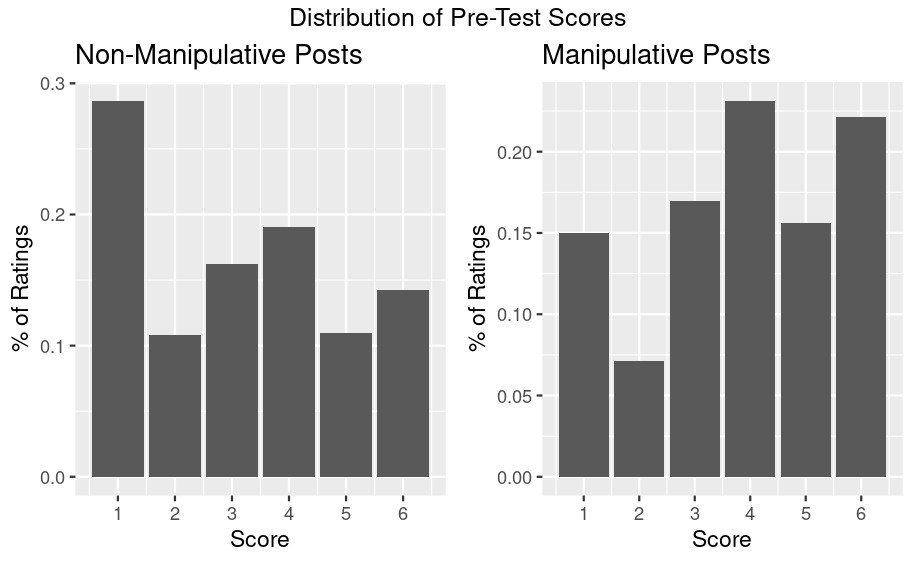
We see that the graph questions result in the largest increase in manipulative answers for both misinfo and true questions (True1 has a valid graph).



#### Deep Dive into Score Distributions

We explore the score distributions of both pre and post-test to gain a better understanding of what answers participants were putting down. We look at the treated group's pre-treatment distribution of answers to non-manipulative questions and manipulative questions.

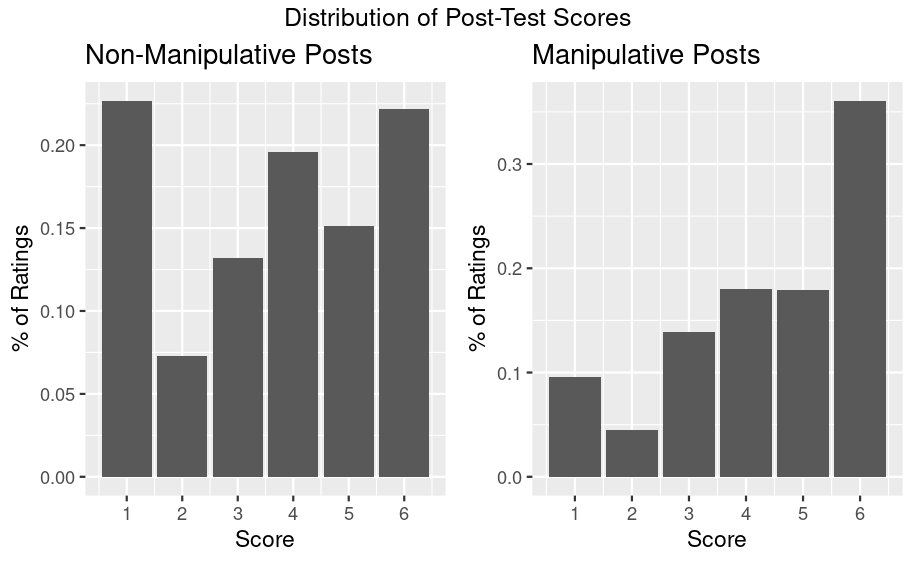
[CODE]



As we can see, pre-test scores for non-manipulative posts are skewed right whereas distribution of pre-test scores for manipulative posts are relatively uniform.

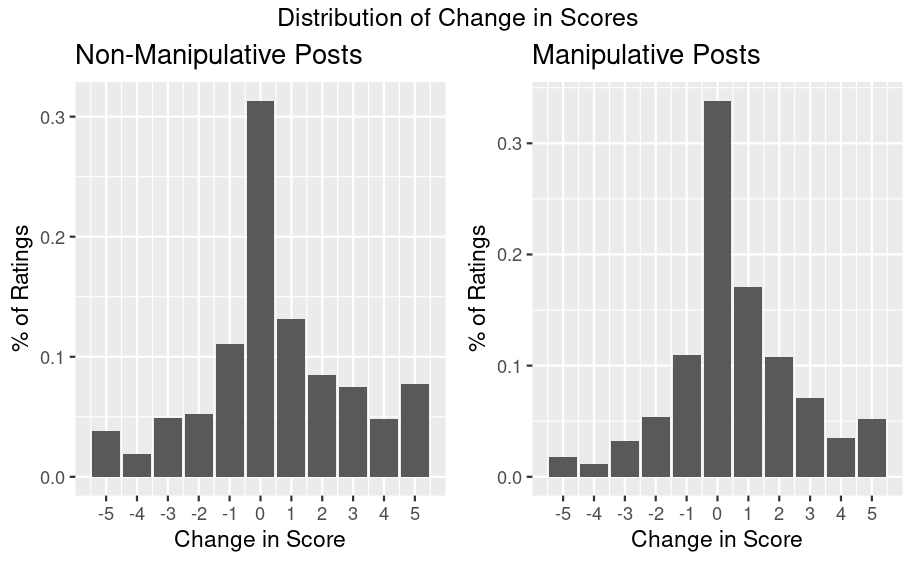
If we look at the Post-Test score distributions, we see that there is a clear shift towards "6 definitely manipulative" for many of the manipulative posts. There is also a shift for non-manipulative posts, and while on average the size of the changes are similar, they are not concentrated at the value 6.

[CODE]



We can also look at the change in score between true and misinfo posts. Both look similar with true having a larger amount of 5 point increases.

Delta



#### Non-Linearity

Although we see a relatively large effect of misidentifying non-manipulative posts, we also suspect that the 6 point scale of labeling for manipulativeness may not be linear. It is easier for participants to change from 2 to 3 than it is for them to change from 5 to 6 since it is easier to be less confident than to be very confident.

To explore this change more formally, we try to binarize the scores to approximate participants' intentions when selecting different scores. We want to understand where would be a suitable cutoff where users really felt a change in manipulativeness.

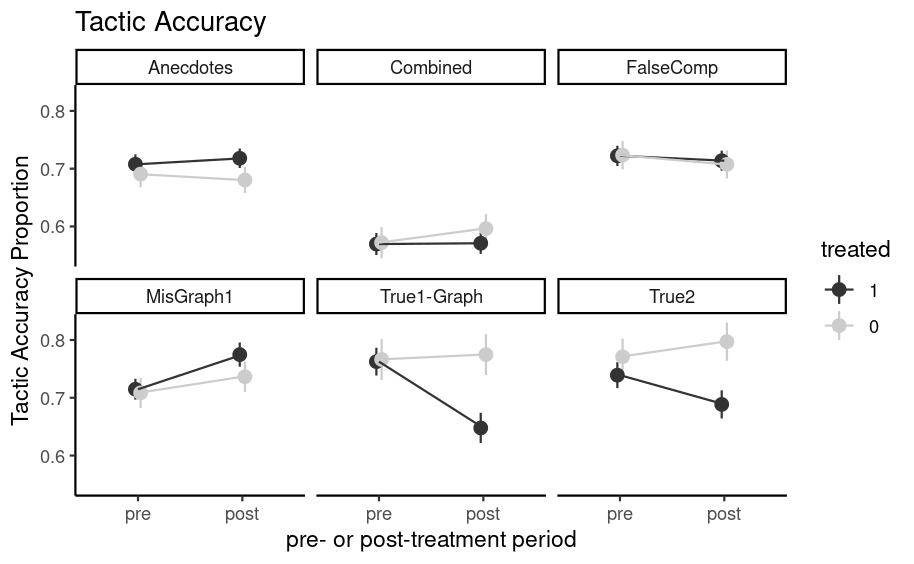
We use the pre-test distributions to explore where on the scale is the greatest relative change in the distribution of answers to the true and manipulative posts. We do this by looking at the difference between the share of answers falling above a potential binary split point for the true versus for the manipulative posts. We calculate these values for all possible binary split points and choose the one where the difference is the greatest. This is where the distributions are the most different in the pre true and manipulative answers. Our goal is to find the best binary split in terms of how people perceive the scale. This method is not ideal as it is still not a direct measure of this perception, however, it is the best we can do given our data.

| **split**  <chr> | **true\_questions\_above**  <dbl> | **false\_questions\_above**  <dbl> | **abs\_diff**  <dbl> |  |
| --- | --- | --- | --- | --- |
| split\_1 | 0.7082495 | 0.8448189 | 0.13656942 |  |
| split\_2 | 0.5995976 | 0.7774145 | 0.17781690 |  |
| split\_3 | 0.4421529 | 0.6081489 | 0.16599598 |  |
| split\_4 | 0.2449698 | 0.3810362 | 0.13606640 |  |
| split\_5 | 0.1388330 | 0.2213280 | 0.08249497 |  |

We see that the optimal split is at a score of 2. Using this split does not change the direction of our insights. Skepticism of both manipulative and non-manipulative posts increased after our treatment.

### Tactics

The following figure plots how accurately participants could identify the manipulative tactics (or lack thereof) in the test questions. These results are investigated in more detail by H2-H4 below.



[High level summary?]

#### 

#### H2 - Misleading Graphs Tactic

H2 is that: participants will be more capable of identifying misleading graphs after taking the course.

As shown in the table below, our results indicate 0.12 points increase in manipulativeness. However, this finding is only significant before adjustments.

The results above are based on a binary outcome variable which measures whether participants pinpointed that the misleading element of each question was based on the graph. If the participant selected the correct misleading element, but also selected another tactic that wasn't present, they would be penalized for the whole question.

Our other outcome measure focuses on each of the three tactics, one at a time. For each individual tactic, for each question participants get a 1 if it is correctly selected when the tactic present, or correctly not selected when tactic not present. Otherwise, they get a 0. Then, we take the average of the six zeroes or ones across all six questions for the given tactic. This percent is the accuracy of their ability to identify the given misleading tactic. (We use the same procedure for all three tactics).

It is also possible to construct an outcome variable which measures the proportion of manipulative tactics that participants correctly identified. For instance, if they selected that a question included a misleading anecdote, when it only included a misleading graph, then they wouldn't score 0, but this mistake would dilute their score. The table below summarizes the results of re-running the analysis using this second outcome variable. As illustrated these results are insignificant before and after adjustment.

#### H3 - Anecdotes Tactic

H3 is that: participants will be more capable of identifying anecdotes after taking the course.

As shown in the table below, participants' ability to identify anecdotes doesn't appear to improve with treatment. This is true for both the binary and proportional outcome variables.

#### H4 - False Comparisons Tactic

H4 is that: participants will be more capable of identifying false comparisons after taking the course.

As shown in the table below, participants' ability to identify false comparisons doesn't appear to improve with treatment. This is true for both the binary and proportional outcome variables.

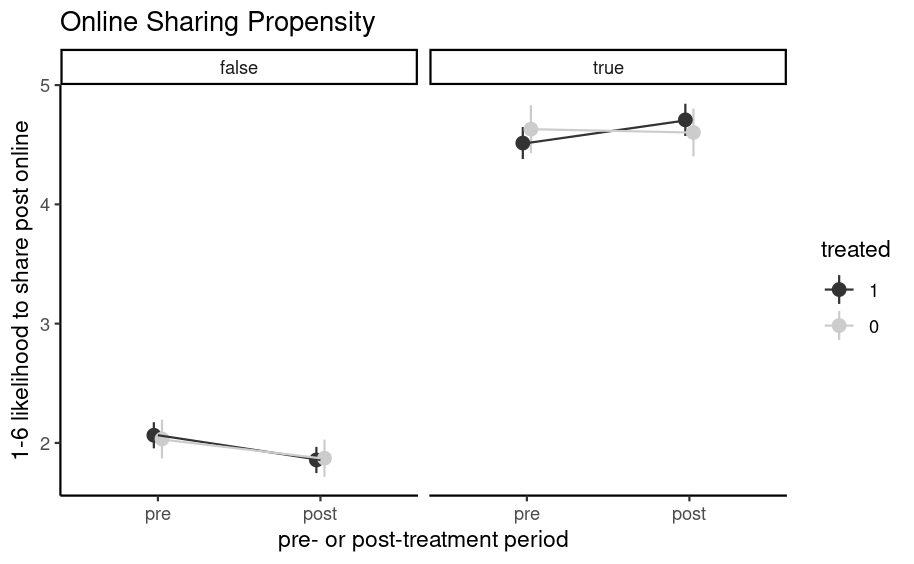
#### Other

[Optional, maybe only include in final report]

### Sharing Actions

#### H5 - Sharing Online

H5 is that: participants will be less likely to share misinformation online after taking the course.



The results below suggest that participants’ propensity to share information online did not change significantly with treatment. This is not entirely unexpected since we did not explicitly steer our treated users towards sharing less through our course. That said, we had hoped that if participants get better at identifying misinformation, then they would share such content less. However, since this secondary effect on sharing would be smaller than the overall effect on identifying misinformation, it is possible that there is a small effect that we did not have enough power to detect.

We may need to design specific lessons in the course to influence sharing directly (e.g. explaining that sharing misinformation could be harmful) to find larger, more easily detectable effect sizes. We also didn't explicitly test to see if some participants may share misinformation with the intention of debunking it. However, since sharing misinformation for any reason is harmful, as it increases its visibility and popularity, it is the overall sharing metric that is the most crucial one and we chose to focus on.

[T Test]

#### H6 - Sharing Offline

H6 is that: participants will be less likely to share misinformation offline after taking the course.

[Graph]

As with H5, the results below suggest that participants’ propensity to share information offline did not change significantly with treatment. Again, this behavior may need to be something that treatment targets explicitly in order to influence it or a much larger sample size.

[T Test]

### Heterogeneous Treatment Effects (HTE)

#### H7 - Susceptibility to Misinformation

H7 is that: participants with different levels of susceptibility to misinformation at baseline will react differently to the treatment in terms of their overall ability to identify manipulative content.

We investigated this hypothesis using three measures of misinformation susceptibility from our survey:

1. Participant performance on the pre-test (`pre\_score`)
2. Participants who self-identified as knowing how to spot manipulative techniques used in articles or headlines (`KnowSpot`)
3. Participants who self-identified as having seen a manipulative news article or headline (`SeenManipulative`)

The results of HTE analysis based on these three variables are shown below. None of the effects are significant.

The results based on `pre-score` may indicate that our test questions were sufficiently neutral to remove any effect. Nevertheless, some prior literature has found HTE based on misinformation susceptibility measured indirectly. Our results based on `KnowSpot` and `SeenManipulative` may indicate that participants struggle to self-identify their own susceptibility (see our correlation graph below).

#### H8 - Political Ideology

H8 is that: participants with different political ideologies will react differently to the treatment in terms of their overall ability to identify manipulative content.

We investigated this hypothesis using a question that asked participants to self-identify as one of "very liberal," "moderately liberal," "moderate," "moderately conservative," and "very conservative." The results, shown below, indicate that there was no significant HTE. This likely reflects the fact that we deliberately chose outcome questions to test misinformation that wouldn't be politically polarizing.

#### H9 - Income

H9 is that: participants with different levels of income will react differently to the treatment in terms of their overall ability to identify manipulative content.

We investigated this hypothesis using a question that asked participants to indicate which range best described their income: less than \$25,000, \$25,000-\$49,999, \$50,000-\$74,999, \$75,000-\$99,999, \$100,000-\$149,999, or \$150,000 or more. We combined these categories to measure participant income as: less than \$25,000, \$25,000-\$49,999, \$50,000-\$99,999, or \$100,000 or more. The results, shown below, indicate that there is a moderately significant HTE. In particular, they suggest that participants with higher incomes had smaller treatment effects. This result is in line with prior literature.

#### New HTE?

#### Causal Trees?

## Controlling Variance

### Covariate Balance

We check whether different covariates that we later use for HTE analysis are balanced between treatment and control groups and correct for any imbalances.

[Code for chart of imbalance]

We find that ‘Race’, ‘Block Users Social Media’, and ‘Report Users Social Media’ are not balanced between treatment and control. Since the blocking and reporting covariate are highly correlated, we decided to only correct for ‘Block Users Social Media’ and ‘Race’ in our new calculations.

#### Correcting Covariate Imbalance

[Regression adjustment code]

Regression adjustment for these covariates largely improved the estimates, and rarely increased standard error for some outcomes but by a very small percentage.

### Correcting for Randomization of Question Order

To control for any variation in the difficulty of the exact matched pre and post test questions we show participants, we randomized which question each person sees in the pre and in the post question, from each of the six pairs of questions.

We also corrected for this randomization and showed minor improvements in the estimates for nearly all outcomes

## Robustness Checks

### Attention Check for Robustness

Using “pass at least 1 attention check” increases our sample size from 641 to 838, which is 84.3% of our total participants. With this larger set, our effect sizes are reduced a little, but not enough to change any of our insights. Since we care about the quality of our answers, we used our “pass both attention check” criteria for our analysis.

### Power Confirmation

With the actual data, we can calculate the power of every question.

[Code]

Our power calculations are based on the summary statistics above. In particular, our power calculations use the standard deviations to calculate post-hoc power with 10 $p$-value adjustments.

Overall, our pre-experiment power calculations were good, but our estimates were just okay. We observe a 6.5% improvement in the main detection task. This is good. However, we knew that this a 6.5% treatment improvement would result in just having marginally enough power (and we see that play out above). Implementing Romano Wolff may possibly help in this regard.

Importantly, we didn’t observe any real difference in the standard deviations between the control and treatment groups for the main task. This was unfortunate: we made this assumption and used imbalanced control and treatment group sizes in the hope of increasing our power. Doing a post-hoc power calculation, our power would have marginally improved if we allocated participants to control and treatment equally.

## Additional Insights

### Time Spent

To better understand how participants interacted with the survey, we analyzed the time they spent engaging with different components. First, we tested whether the control and treatment groups spent a different amount of time completing their respective course.s The difference is significant, but only before adjustments.

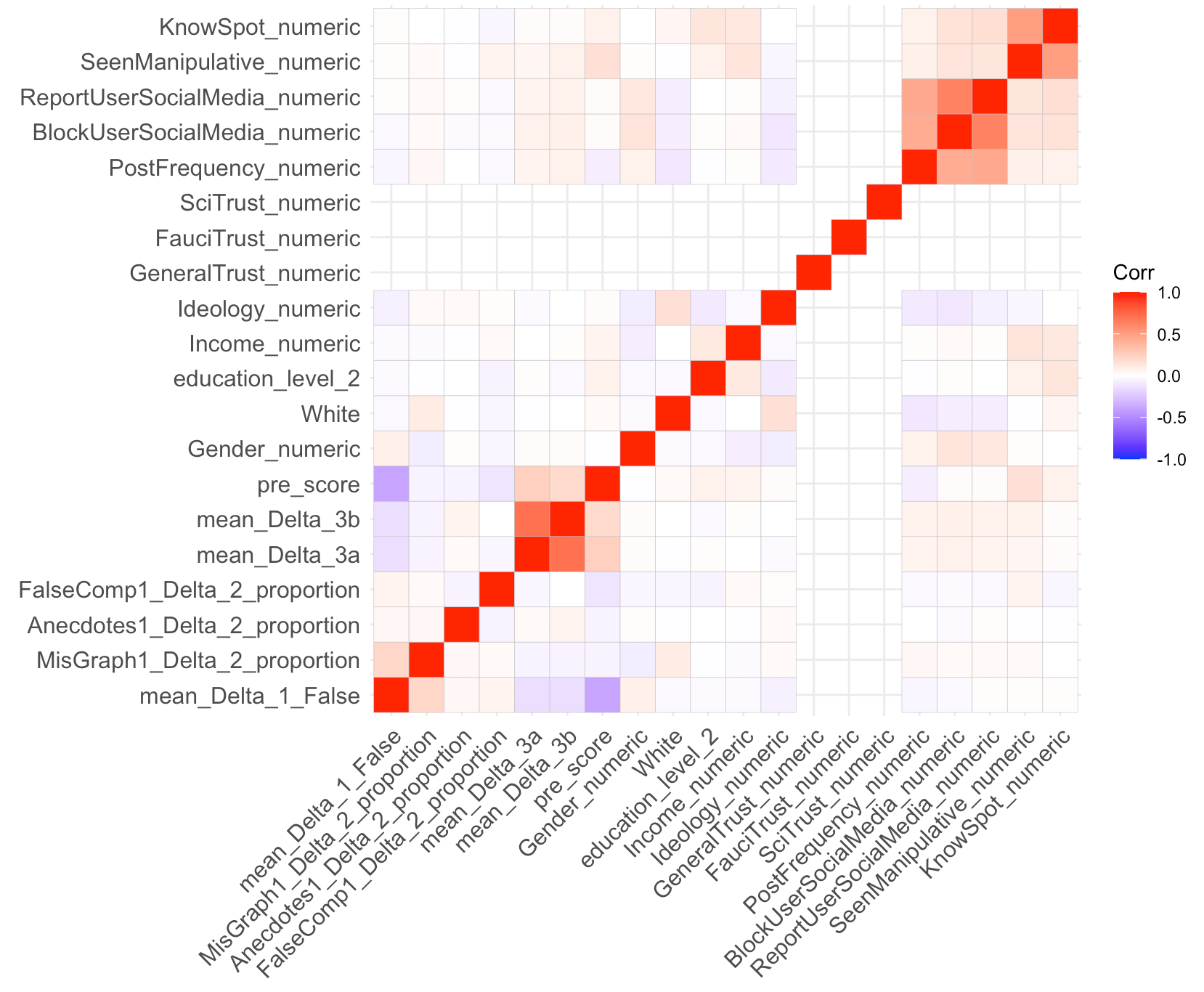
Next, we tested whether the time survey participants spent on the test pre-treatment was different to the test post-treatment. Note that, for this test, we included all participants regardless of whether they were in the treatment or control arm. The results suggest that the difference is significant, both before and after treatment. In particular, participants spent, on average, less time post-treatment than they did pre-treatment.

Next, we tested whether this same difference between the pre and post-tests was present if we only consider treated participants. The results suggest that the difference is significant, both before and after adjustments.

Finally, we tested if the different time participants spent on the pre and post-tests differed based on their allocation to the control and treatment arms. The results suggest that the difference wasn't significant, even before adjustments.

### Pretreatment Correlations

[Graph Code]



We see that only pre-score (i.e. baseline susceptibility score calculated from pre-test survey responses) is correlated with the main outcomes of the experiment. This makes sense as the main outcomes are calculated as post-score - pre-score, i.e. we have already adjusted for pre-score in our design of the experiment. It is interesting to note that pre-score is also correlated with the sharing behavior measured by mean\_Delta\_3a and mean\_Delta\_3b. Correcting for the pre-score covariate did not change the estimates for sharing outcomes by a lot.

[CODE CHUNKs WITH RESULTS]

* Original susceptibility
* Pre sharing

### ~~Participant Comments~~

~~[Optional, maybe only include in final report]~~

## Future Experiments

Our experiment could only yield limited insights due to budget and time constraints. However, these insights prompt additional questions and provide directions for future research.

In particular, we suggest 3 directions of additional research on this topic.

1. Longer term effects of treatment (and in general longer time for delivering treatment messages)
2. Behavior focused treatments, focusing on explaining the importance of not sharing misinformation
3. Measuring misinformation identification with evidence search or with observations on real world behavior