# Project Analysis Plan

## Research Questions

Our experiment will focus 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

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.

## Power Confirmation

Using actual results, report all components related to power.

1. What is our usable sample size (passing different tiers of attention checks)?
2. What is our effect size?
3. What is our SD?
4. What is our power?

| **treated**  <chr> | **n**  <dbl> | **Delta\_1\_False\_power**  <dbl> | **Delta\_1\_True\_power**  <dbl> | **Delta\_2\_power**  <dbl> | **Delta\_3a\_power**  <dbl> | **Delta\_3b\_power**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 214 | 1 | 0.99 | 0.1 | 0.01 | 0.04 |
| 1 | 427 | 1 | 0.99 | 0.1 | 0.01 | 0.04 |

| **treated**  <chr> | **n**  <int> | **Delta\_1\_False\_mean**  <dbl> | **Delta\_1\_False\_sd**  <dbl> | **Delta\_1\_True\_mean**  <dbl> | **Delta\_1\_True\_sd**  <dbl> | **Delta\_2\_mean**  <dbl> | **Delta\_2\_sd**  <dbl> |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 214 | 0.1133178 | 1.132510 | 0.1588785 | 1.676401 | 0.08878505 | 1.306023 |  |
| 1 | 427 | 0.7394614 | 1.160356 | -0.6147541 | 1.842621 | -0.10538642 | 1.389310 |  |

| **treated**  <chr> | **n**  <int> | **Delta\_1\_False\_mean**  <dbl> | **Delta\_1\_False\_sd**  <dbl> | **Delta\_1\_True\_mean**  <dbl> | **Delta\_1\_True\_sd**  <dbl> | **Delta\_2\_mean**  <dbl> | **Delta\_2\_sd**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 214 | 0.1133178 | 1.132510 | 0.1588785 | 1.676401 | 0.08878505 | 1.306023 |
| 1 | 427 | 0.7394614 | 1.160356 | -0.6147541 | 1.842621 | -0.10538642 | 1.389310 |

| **treated**  <chr> | **n**  <int> | **Delta\_3a\_False\_mean**  <dbl> | **Delta\_3a\_False\_sd**  <dbl> | **Delta\_3b\_False\_mean**  <dbl> | **Delta\_3b\_False\_sd**  <dbl> |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 214 | -0.161215 | 0.7929800 | -0.2242991 | 0.9366273 |  |
| 1 | 427 | -0.207260 | 0.9550121 | -0.3079625 | 1.0083314 |  |

How do all these compare to our estimates? Were there items that were dramatically different from our predictions? Why was this the case and does it affect our confidence in the results?

In the table the Delta columns are the mean deltas for a given question and SDs the given standard deviations. Standard deviations are what we used for the power calculations. Then, the power columns represent post-hoc power calculations with 10 p-value adjustments.

Overall, our power calculations were good, but our estimates were just okay. We observe a 6.5% improvement in the main detection task. That’s good! However, we knew that this a 6.5% treatment improvement would result in just being on the edge in terms of power, and we see that playout. [As show below] The main task is not significant after doing a BH correction but is significant before corrections. Possibly, implementing Romano Wolff may help in this regard.

Even still, the confidence interval is wide, with the confidence interval ranging from an improvement of .4 to 12.6%.

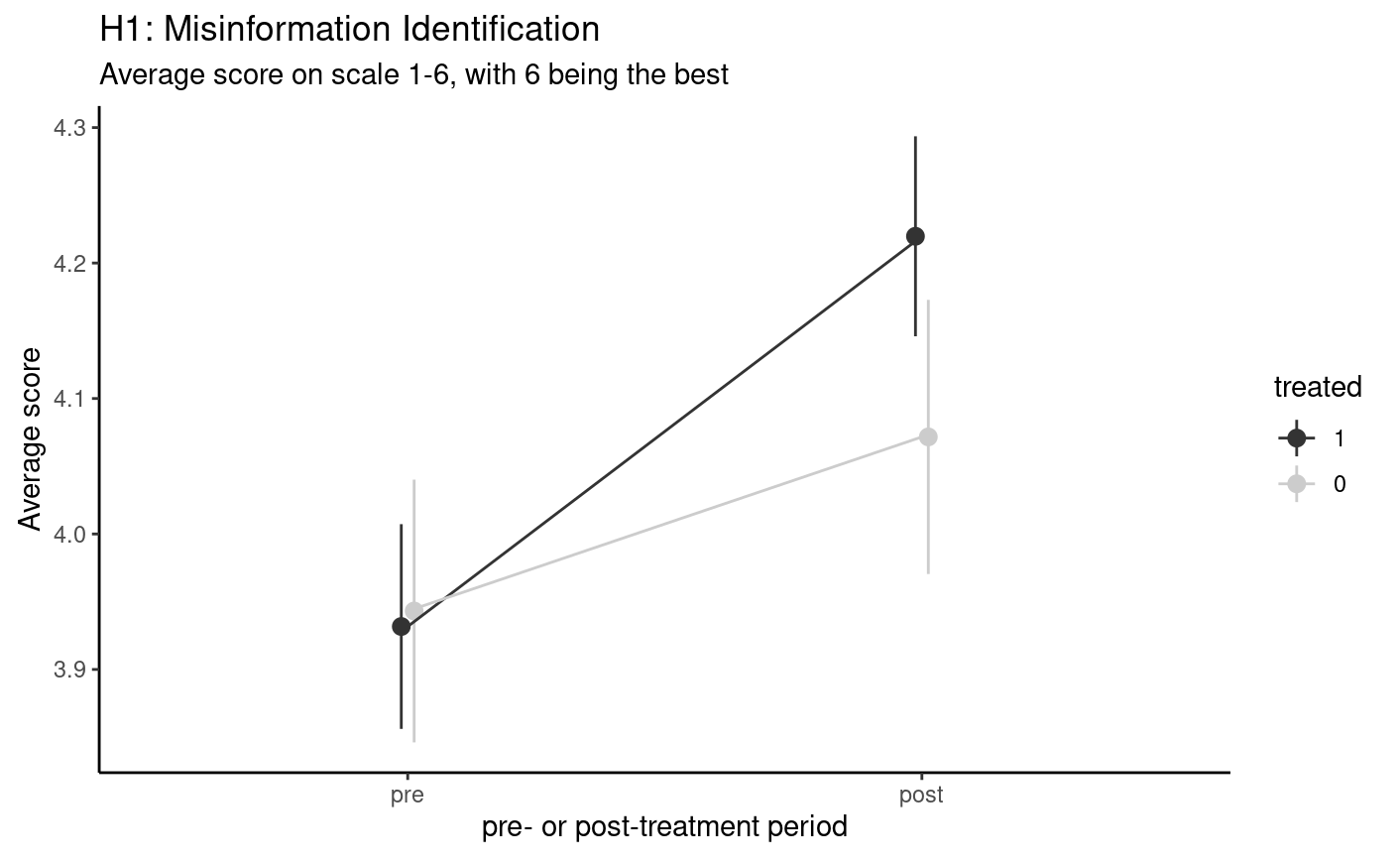
One thing we didn’t see is any real difference in the standard deviations between the two groups for the main task. Doing a post-hoc power calculation, our power would have improved marginally to .29 (instead of .26) if we used equal groups.

## Treatment Effect of Identifying Manipulation

What is the overall impact of our experiment? Stat sig? (H1)

Combined (all questions)

| **estimate**  <dbl> | **estimate1**  <dbl> | **estimate2**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> | **parameter**  <dbl> | **conf.low**  <dbl> | **conf.high**  <dbl> | **p.value\_adjusted**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.16 | 0.29 | 0.13 | 2.15 | 0.03 | 419.96 | 0.01 | 0.31 | 0.32 |



Here is the output of the standard t.test function in R. The estimate column represents the delta in the group means (treated versus control, same delta if you subtracted the two groups in the above table). 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. Conf.low and conf.high are the confidence intervals. P.value\_ajusted is the **adjusted** p-value with 10 adjustments using the BH correction.

We do see an improvement of 23% (.16/.13 - 1) which is significant before adjustments but is not significant after adjustments. Could we possible consider fewer adjustments?

Note that out of our 6 point scale, .16 is 3.2% (.16/5). This means combined treatment moved the results around 3.2% more on our scale (3.2% more manipulative?).

To better understand the results above, we segment the questions between true and manipulative questions.

(False questions only)

| **estimate**  <dbl> | **estimate1**  <dbl> | **estimate2**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> | **parameter**  <dbl> | **conf.low**  <dbl> | **conf.high**  <dbl> | **p.value\_adjusted**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.63 | 0.74 | 0.11 | 6.55 | 0 | 435.77 | 0.44 | 0.81 | 0 |

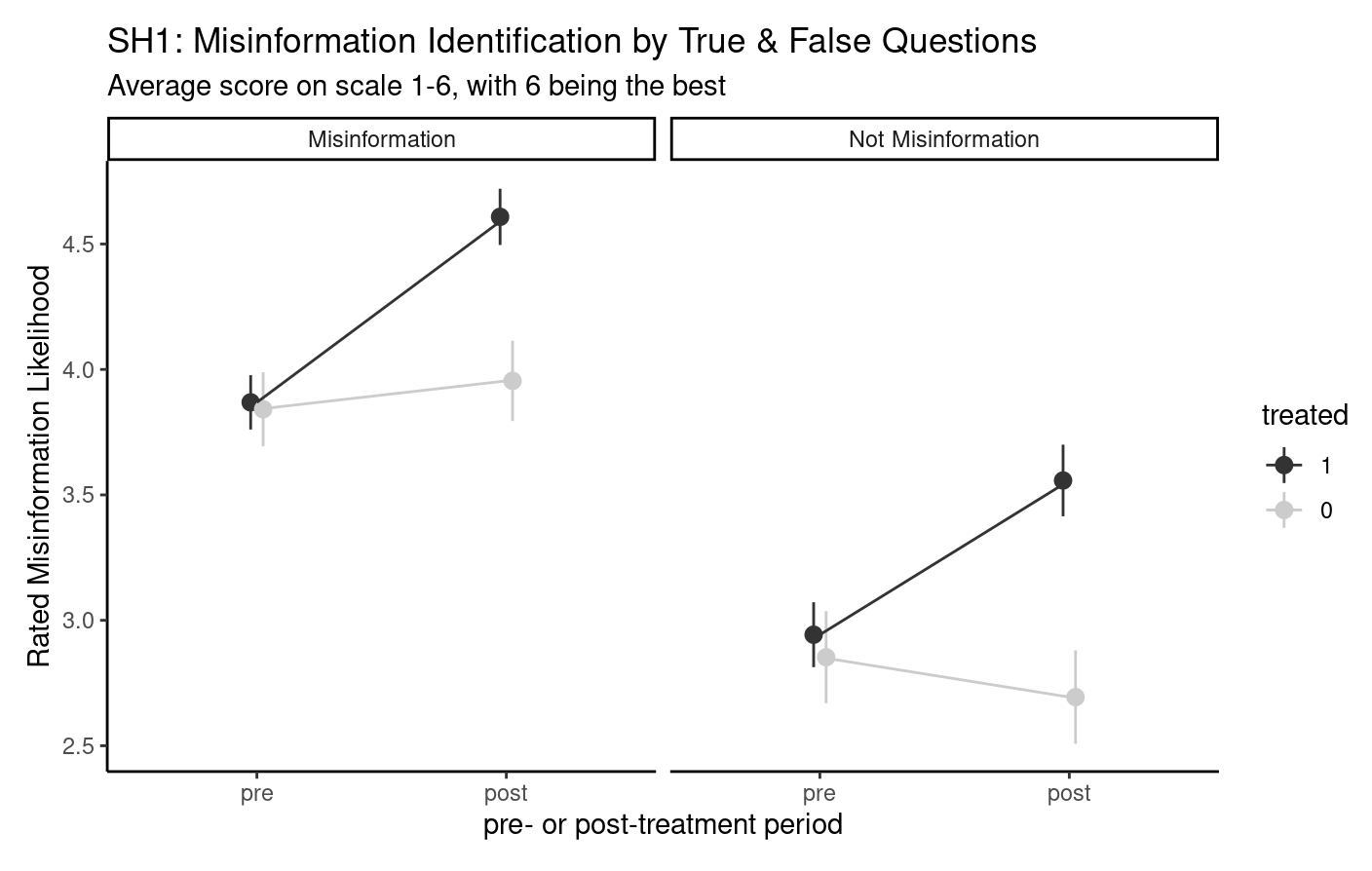
(true questions only)

| **estimate**  <dbl> | **estimate1**  <dbl> | **estimate2**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> | **parameter**  <dbl> | **conf.low**  <dbl> | **conf.high**  <dbl> | **p.value\_adjusted**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| -0.77 | -0.61 | 0.16 | -5.33 | 0 | 463.97 | -1.06 | -0.49 | 0 |

We see that we have much larger effect sizes. For false questions, treated people found questions to be 12.6 percentage points (.63/5) more manipulative or 472% (.63/.11-1) increase in manipulativeness.

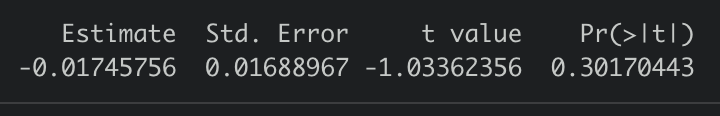
However, we also see that our treatment caused people to be more suspicious of true questions. True questions are flipped, so negative means answering the question more wrong. This effect might not be as long lasting as the knowledge they gained from learning to identify misinfo.

We’re also assuming linear change in scores right now. However, it could be harder to go from 5 to 6 for manipulativeness than it is to go from 1 to 2 for trues. We need to figure out how to adjust for this.



HTE (H7-9)

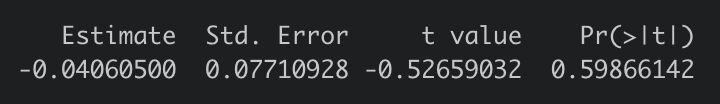
H7: Participants with different levels of susceptibility to misinformation at baseline will react differently to the treatment in terms of their overall ability to identify manipulative content.



We did not find HTE based on the baseline susceptibility to identify misinformation. The above table shows the estimate of interaction effect of baseline score and treatment on our main outcome (ability to identify misinformation).

These results are a bit surprising as we expected participants with smaller baseline scores to show a larger post - pre score. the short duration of the course gives a possible explanation for these results. We can expect longer duration courses to cause statistically significant differences for participants with smaller baseline scores to account for time lag in inculcating treatment information.

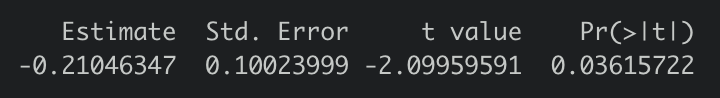
H8: Participants with different political ideologies will react differently to the treatment in terms of their overall ability to identify manipulative content.



We did not find HTE based on political ideology. The above table shows the estimate of interaction effect of political ideology and treatment on our main outcome (ability to identify misinformation).

These results make sense as the examples we used in pre and post treatment surveys were neutral. We did not want to use polarizing examples that could incite strong emotions and influence the decision making ability of participants.

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

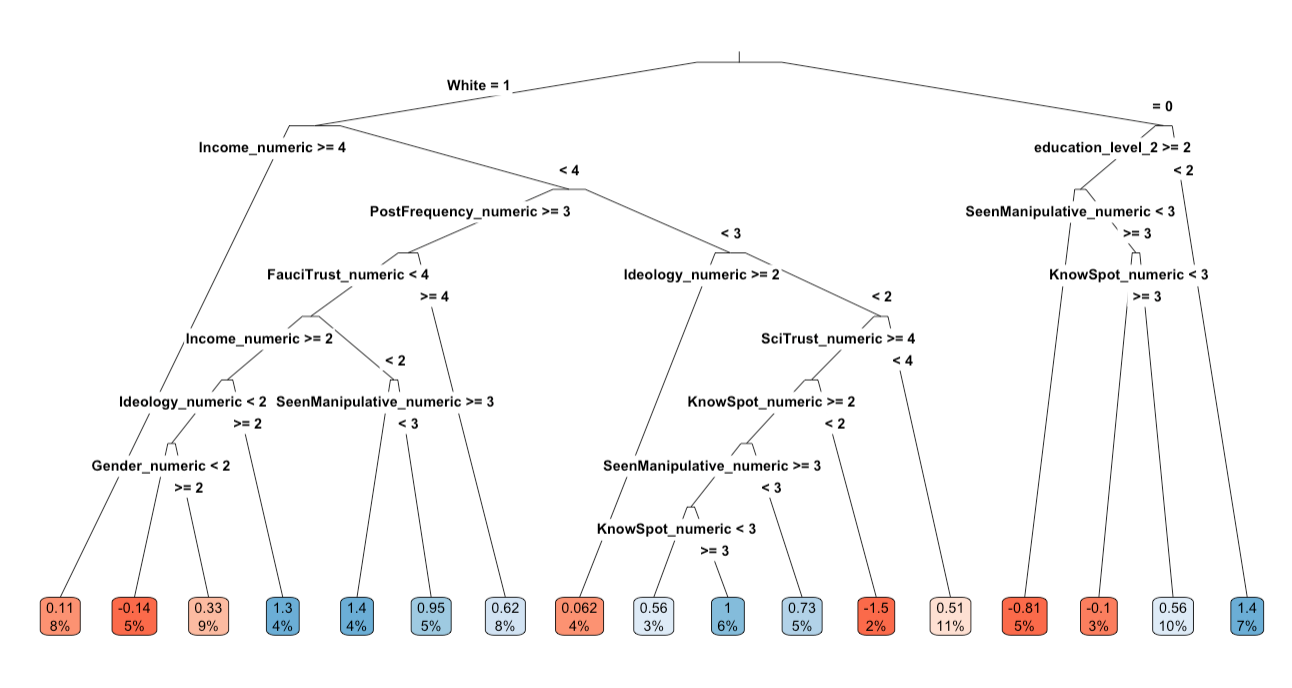


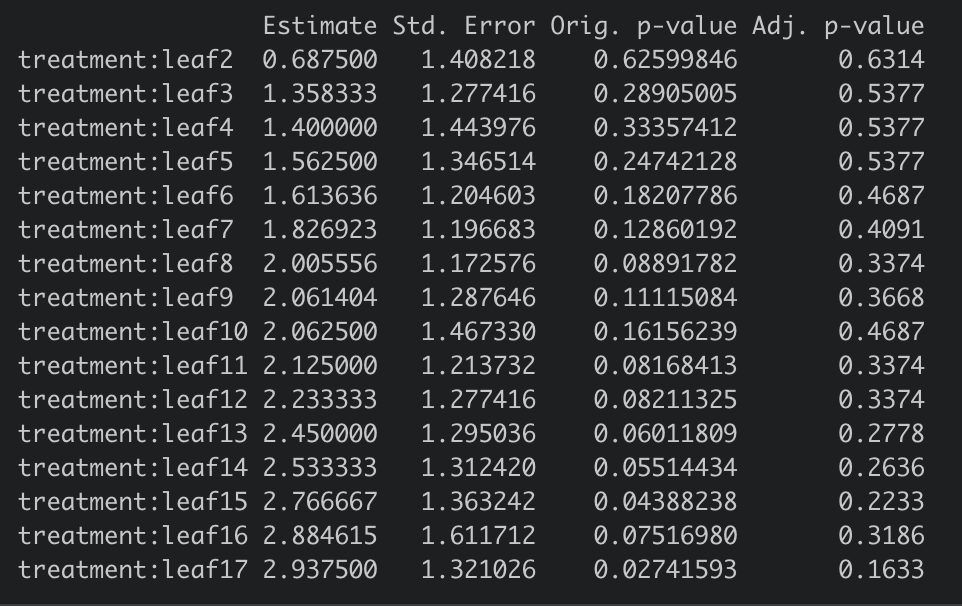
We found HTE based on income levels. The above table shows the estimate of interaction effect of income levels and treatment on our main outcome (ability to identify misinformation).

Income had prominent HTEs in the pilot as well, and continues to show HTE in our main survey.

Casting doubt, does the treatment cause people to misidentify true questions? (SH1)

**Causal Trees**

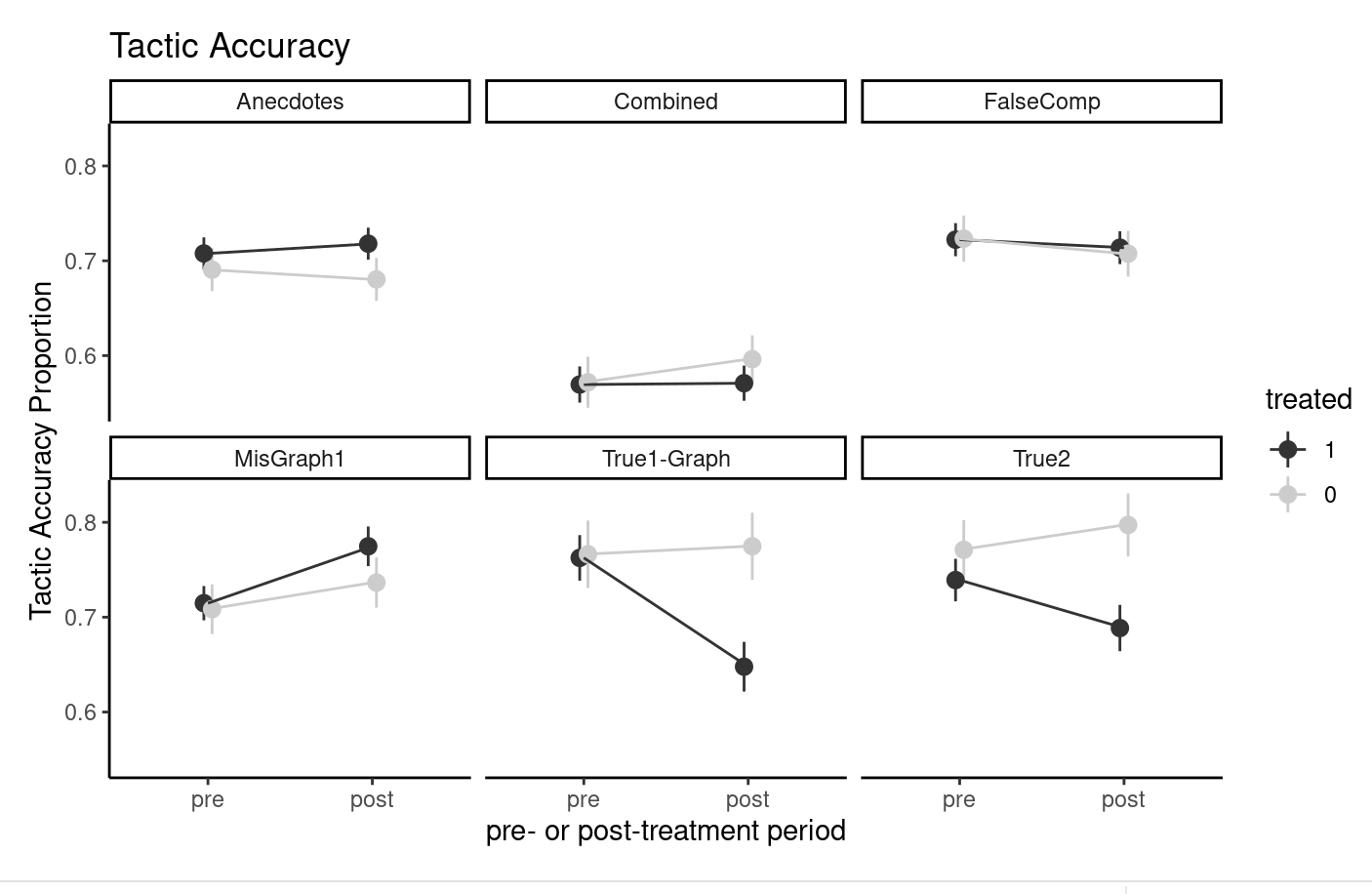
****

****

## Treatment Effect on Specific Tactics

Looking at the test questions separately (correlation/baseline/check):

Are some questions harder than others? (based on pre-test)



Did people learn some tactics more easily than others? (based on comparing treatment effect size per question) (H2-4)

Misleading Graph



Here is the output of the standard t.test function in R. The estimate column represents the delta in the group means (treated versus control, same delta if you subtracted the two groups in the above table). 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. Conf.low and conf.high are the confidence intervals. P.value\_ajusted is the **adjusted** p-value with 10 adjustments using the BH correction.

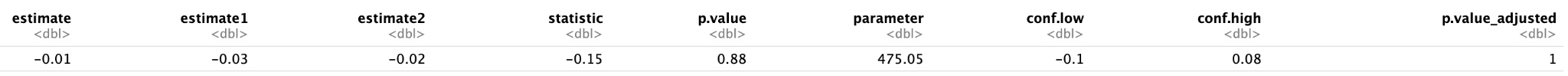
People’s detection of misleading graphs improved by quite a bit. Similar to H1, the p-value is significant before adjustments.

Anecdotes



People’s detection of anecdotes doesn’t appear to improve with the treatment.

False Comparison



People’s detection of false comparison doesn’t appear to improve with the treatment.

I think even without the significant result after adjustment for misleading graphs, these results suggest that more testing should be done on First Draft’s ability to improve people’s ability to identify misleading graphs.

Looking at the tactics question:

Specifically identifying the tactic (analyzing the tactics questions - correctly selecting or not selecting given tactic across all 6 questions)(H2-H4)

## Treatment Effect on Sharing

Online/Offline, does our treatment reduce desire to share or discuss? (overall) (H5, H6)

H5

Propensity to share misinfo (only misinfo questions)

| **estimate**  <dbl> | **estimate1**  <dbl> | **estimate2**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> | **parameter**  <dbl> | **conf.low**  <dbl> | **conf.high**  <dbl> | **p.value\_adjusted**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| -0.05 | -0.21 | -0.16 | -0.65 | 0.52 | 502.46 | -0.19 | 0.09 | 1 |

1 row

Individuals’ propensity to share online did not change significantly. This makes sense since we did not explicitly steer our treatment users towards reducing sharing through our course. We would need to design specific lessons in the course to influence sharing directly (eg explaining how sharing misinfo is harmful).

We also took out the option of “sharing to inform”.

H6

Propensity to share misinfo offline (only misinfo questions)

| **estimate**  <dbl> | **estimate1**  <dbl> | **estimate2**  <dbl> | **statistic**  <dbl> | **p.value**  <dbl> | **parameter**  <dbl> | **conf.low**  <dbl> | **conf.high**  <dbl> | **p.value\_adjusted**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| -0.08 | -0.31 | -0.22 | -1.04 | 0.3 | 455.47 | -0.24 | 0.07 | 1 |

Once again, not significant for sharing offline

Split by topic? Are some topics less likely to be shared? (correlation/baseline/check)

Secondary outcome:

* effect on sharing by tactics
* casting doubt, does the treatment cause people to share less true questions? (SH1)

## Rating Quality/Time Analysis

Can we identify contradictions that people tend to make? Self evaluation of skill vs actual performance

Time spent on post questions as outcome

Does spending longer on questions correlate with better understanding? (baseline/check/correlation; secondary outcome/correlation)

Does spending longer on the course correlate with better outcomes? (baseline/check/correlation; secondary outcome/correlation)

Are there any significant differences if we include both attention checks vs just one?

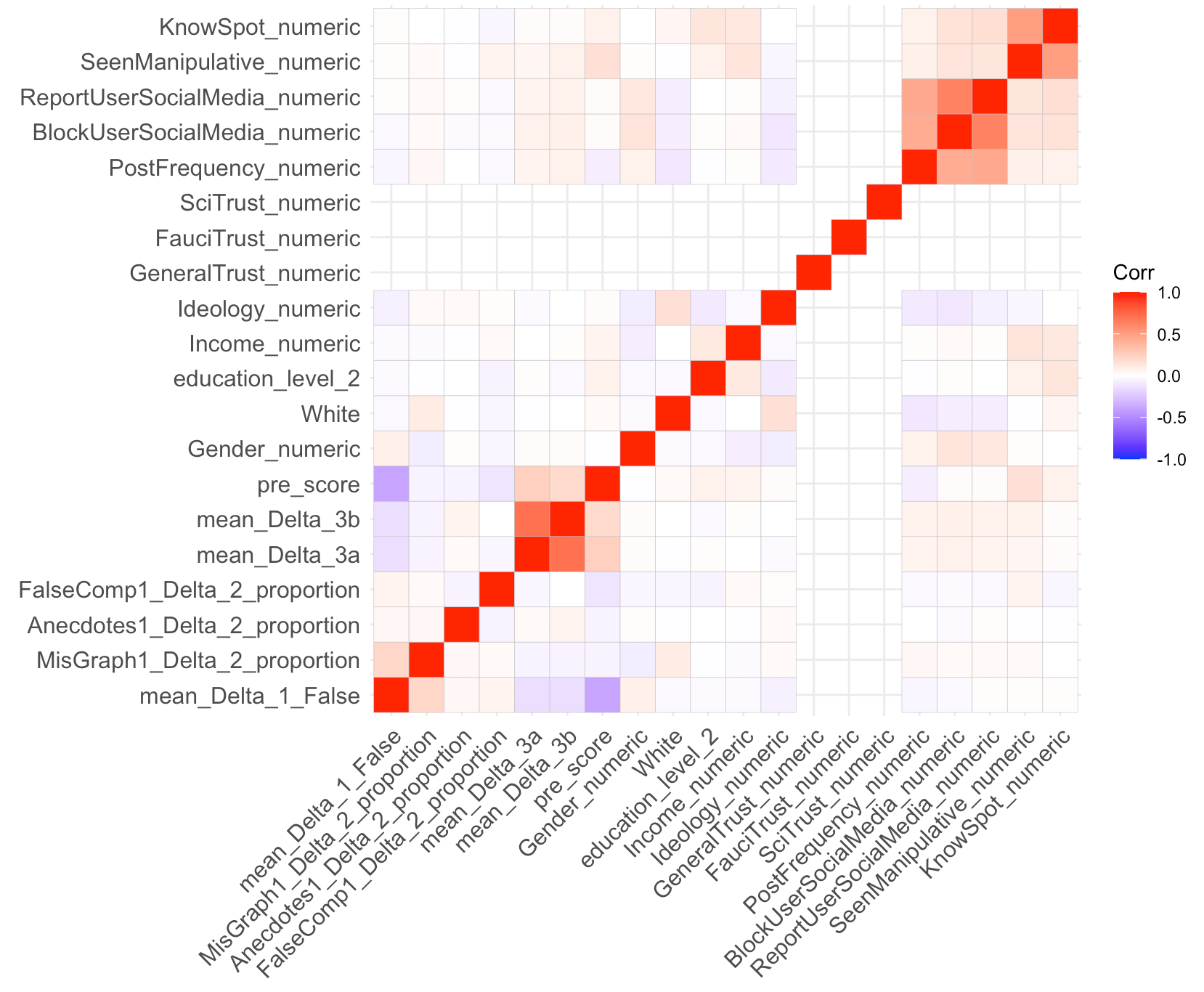
Using “pass at least 1 attention check” reduces our effect sizes by a little bit, but does not change any of our insights. It gives us more power, so maybe we should just use more samples even though they may be slightly lesser quality.

## 

### Baseline/checks/correlations

Pretreatment distributions -

Correlations between outcomes and covariates at baseline.



#### 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 behaviour 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].

#### Balance checks/key covariates

Based on the splits used for HTE, we found 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. Regression adjustment for these covariates largely improved the estimates, and rarely increased standard error for some outcomes but by a very small percentage. We can see the results of covariate correction below [CODE CHUNKs WITH RESULTS].

#### Other important baselines correlations to check/show

We also tried correcting for randomization of the order of questions in the survey. This correction showed minor improvements in the estimates for nearly all outcomes [CODE CHUNKs WITH RESULTS].

### Main outcomes/analysis

#### Outcomes

* Identify manipulative or not correctly (score 1-6)
* Changing in identifying manipulative score post-pre
* Sharing variables: look at non true questions and transform the variables to get score 7 minus raw rating value (1-6)
* Tactic score (sum up correct or not)

#### Key regressors of interest

#### Important covariates

* Attention\_miss (0 is good) (could do with 1) (or ATT1)

#### HTE variables

* Pre ability to identify manipulative content score
* Ideology
* Income

### Other analysis

Preprocessing Tasks

* Pre and post column (combining a and b randomization)
* Total timer for test (by summing across columns)

## Milestone 2 Analysis

Outline

* Intro/Research Questions
* Summary of insights
* Data Processing
* Controlling Variance
  + Controlling for question ordering
  + Covariate balance
* Manipulativeness
  + Increase in manipulativeness
  + True also increase
  + Pretreatment distribution of misinfo is quite uniform compared to True questions which tend to be skewed right
  + Binary interpretation?
* Tactics
  + Exact answer
  + Proportional
  + Other
* Sharing actions
* HTE
  + ML for HTE
* Additional comments
* Attention Check for robustness
* Power Confirmation
* Timers
* Pretreatment correlations
  + Original susceptibility
  + Pre sharing
* Future experiments

**Main Analysis:**

* Clean up sharing power calculation + splitting it between true vs misinfo questions (Quentin)
* Add in graphs/data about H1 split between true and false (H1 + SH1) (Ross)
* Change tactics calculation from exact binary to proportion (Ross)
* HTE on H1 (KG)
* Re-run power simulation for new correction factor + power for individual tactics (KG)
* Preliminary graphs (Ross)
* Explanations (Ben and Quentin)
* Prepare timer columns to be processable (Ben)
* Look at graph true for mistakes (Quentin)

**Secondary Analysis**:

* Details of original hypotheses/analysis
  + Misleading graph, are there specific graphs that people thought were manipulative? (Ross)
  + Add more visualizations for each hypothesis (Ross)
    - Graph by question
    - Graphs for sharing behavior
  + pretreatment distributions, how is sharing behavior before? how is susceptibility? (correlations at baseline - baseline misinfo, sharing, tactic variables and various baseline demographics) (KG)
  + Increasing power for the tactics identification & sharing (KG) (see details below for de-meaning by question order)
* Qualitative checks (check ‘other’ responses) (KG)
  + Other tactics
  + Extra comments
* Are the covariates balanced? (KG)
* Playing with time (Ben)
  + Time spent on control course vs treatment course
  + Time on pre and post test (do they take longer after going through courses)
  + Time on pre and post conditional on control vs treatment (does treatment take more time on evaluating questions after treatment)
* ML for HTE (KG)
* Pretest distribution for how people are responding on the scale (Réka, Quentin)
* Amazon gift card picking (att3 = 0, pick random codes, check email) (Quentin)
  + Email address was [experimentalp281@gmail.com](mailto:experimentalp281@gmail.com), password should be *Econ281class*
* Improve on the narrative for milestone 2 (need another google doc)

**Nice to Know:**

* Pre-test breakdown by demographics (do some types of people share more than others? Are they more accurate/good at identifying misinfo?)
* Identify borderline results for which we need more samples (and how many more?)
* Explore HTE on other covariates too

**Qs to ask Susan**

* Linear scale might not be accurate of reality. Can we assume normal distribution (for some questions and readjust to get a sense of how skeptical we made people for true?
* How do we actually interpret our results, can we state course is effective in the way we thought it was instead of just some priming effect?
* Short summary of main results.
  + Manipulative increase
  + tactics not super accurate except graphs (not sig)
  + sharing no change
  + no HTE
* Balancing the overall increase in skepticism in true and false.
* DIscuss future directions with her.
  + Study of lasting effects
  + Course on influencing behaviors
    - How to ask questions that are reflective of actual actions they will take?
  + Different measure

**Talking with Susan**

* Finding something interesting with$1000 is a big accomplishment
* How will you use this pilot for another pilot?
* Imagine you’ll make a decision at the end of this
* Would do multiple pilots if you were to spend 50k
* Final presentation: Said to do X, did X, results, expected, what we got; what would we do if we had more money to spend (talk a bit about this); point out what we would have done differently; things we didn’t know and would have done differently
* Value judgment on sharing, possibly teaching explicitly about that
* Alternative outlet on sharing
* Short term outcomes
* **Feedback from Tom**
  + Likes visualization
  + Right track
  + Interpreting the scale
    - Pre-test distribution - realistic data
  + Question fixed effect and controlling for other variables that may explain the tactics outcomes
    - HERE
  + Find binary split based on baseline distributions (what looks reasonable)

**Future Directions**

* Lagging effect of treatment? Does the skepticism for true wear off quickly whereas tactics remain remembered?
  + What happens when we have a longer course with more spread out evaluation periods?
  + How do these results apply to more polarizing or other types of misinfo examples?
* New treatments targeting behaviors
  + Do they fact check more or have improved misinfo identification accuracy for an evidence searching task?
  + Can we build sharing information into our treatment course? (we didn’t explicitly tell them sharing misinfo is bad)
* Alternative (better) ways of checking misinformation (like evidence search. etc)
* How do we teach misinformation identification without making people paranoid about everything :P
* Incentivize