# **Pre-Analysis Plan - First Draft**

Reka Zempleni, Ben Ditchfield, Quentin Hsu, Ross Dahlke, Kushagra Gupta

April 2022

## **Instructions:**

This is a template of an experimental plan document, or what is called a Pre-analysis Plan (PAP). In this course, your team will work together on producing a PAP for your experiment. It will be one of the final deliverables.

The PAP is a planning tool, to help your team structure your research from design to analysis, think through all the necessary steps before implementation, lay out specific hypotheses you want to test, and plan for analysis before the data comes in. This can potentially lower risks of ex-post cherry-picking of causal relations.

This template is produced to help you frame your planning of the experiment. With resources including the Experimental Project Packet, discussions with partners and experts, and the experimental design tutorials, you should be able to complete the various components necessary for a PAP.

This tutorial was originally developed for the Spring 2021 course, ALP301 Data-Driven Impact.

## **Motivation**

*Please briefly explain the motivation for your project, objectives, and what are the implications of the study.*

Online misinformation has become a crucial threat to a well-informed citizenry. False/manipulative statements and conspiracy theories are widely circulated on social media, challenging our public health responses to the pandemic and eroding voters’ trust in democratic institutions. This project seeks to identify a low-cost way of increasing people’s ability to recognize online misinformation. In particular, it focuses on teaching people to recognize manipulative tactics in online information, and reduce their likelihood of sharing manipulative content online and offline.

First Draft has created a “Protection from deception” SMS course that teaches users to recognize misinformation through a series of text messages over 14 days. In principle, our experiment is intended to test the efficacy of such a course. But it has also been designed to navigate several constraints:

* The experiment is limited to a 15-minute online survey. This survey is responsible for collecting participant demographics/covariates, delivering the treatment or control messaging, and evaluating its effectiveness by implementing a pre and post-treatment test. This time-limit constrains the length and depth of the course we can show to participants.
* The online survey is unable to target participants based on their characteristics (such demographics or misinformation susceptibility). This prevents customizing the content for a particular audience. The survey must instead be accessible by and effective for a representative sample of the US population.
* The study has a limited sample size of 1,000-1,5000 participants. First Draft also indicated that they expect the treatment effect to be small.

Taking these together, we are wary of the experiment’s power. To navigate these constraints, we considered results from the pilot, the literature, and input from First Draft. We’ve adapted and supplemented a subset of First Draft’s SMS course to serve as our treatment. Our 6-minute “mini-course” includes a mixture of new and existing content focusing on specific manipulative tactics. First Draft requested we include new content previously untested in the existing literature.

Our course teaches participants to recognize three features common to misinformation based on a taxonomy developed in Cook (2020): misleading graphs, anecdotes, and false comparisons (the latter two being novel content). Our approach focuses on teaching participants to recognize tactics rather than misinformation itself. This is intentional. Audiences often require significant context, topic-specific knowledge, or access to evidence to correctly identify misinformation. Manipulative tactics, on the other hand, are topic-agnostic and objective. We can therefore expect participants to more accurately identify these tactics regardless of their existing knowledge. The **Treatment** section further elaborates on the rationale for our design choices.

Our experiment will investigate whether participants are able to better recognize these tactics by the end of the course relative to the control group. We will also test the effects for each tactic independently. Furthermore we will test the effects of our treatment on follow-up actions such as sharing the information on social media and discussing it offline. Finally, we will explore heterogeneous treatment effects (HTE) based on two covariates we collect (income and ideology) and the participants’ performance on the pre-intervention test.

## **Research Questions**

*What are the research questions that your experiment tries to address?*

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 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?

## **Experimental Setup**

### **Sample Recruitment**

*How, where and when would your participants be recruited? Who are your participants? Why is this an appropriate sample? How is the data going to be collected?*

Participants will be adults 18 years and older living in the US, recruited through Lucid. We target a nationally representative sample to obtain more generalizable results. The original intention of First Draft’s SMS course is to educate US citizens on election-related disinformation. Using this representative sample makes our results generalizable and useful for First Draft.

The data will be collected through a self-paced online survey implemented via Qualtrics.

Our budget for the experiment is 1,000-1,500 participants. We expect there to be some drop-off from attention checks, so we used a conservative sample size target of 800 for our pre-analysis calculations.

### **Experimental Design**

*Please explain your experimental design. For example, is the experiment looking at within participant differences or between participant differences in outcomes?*

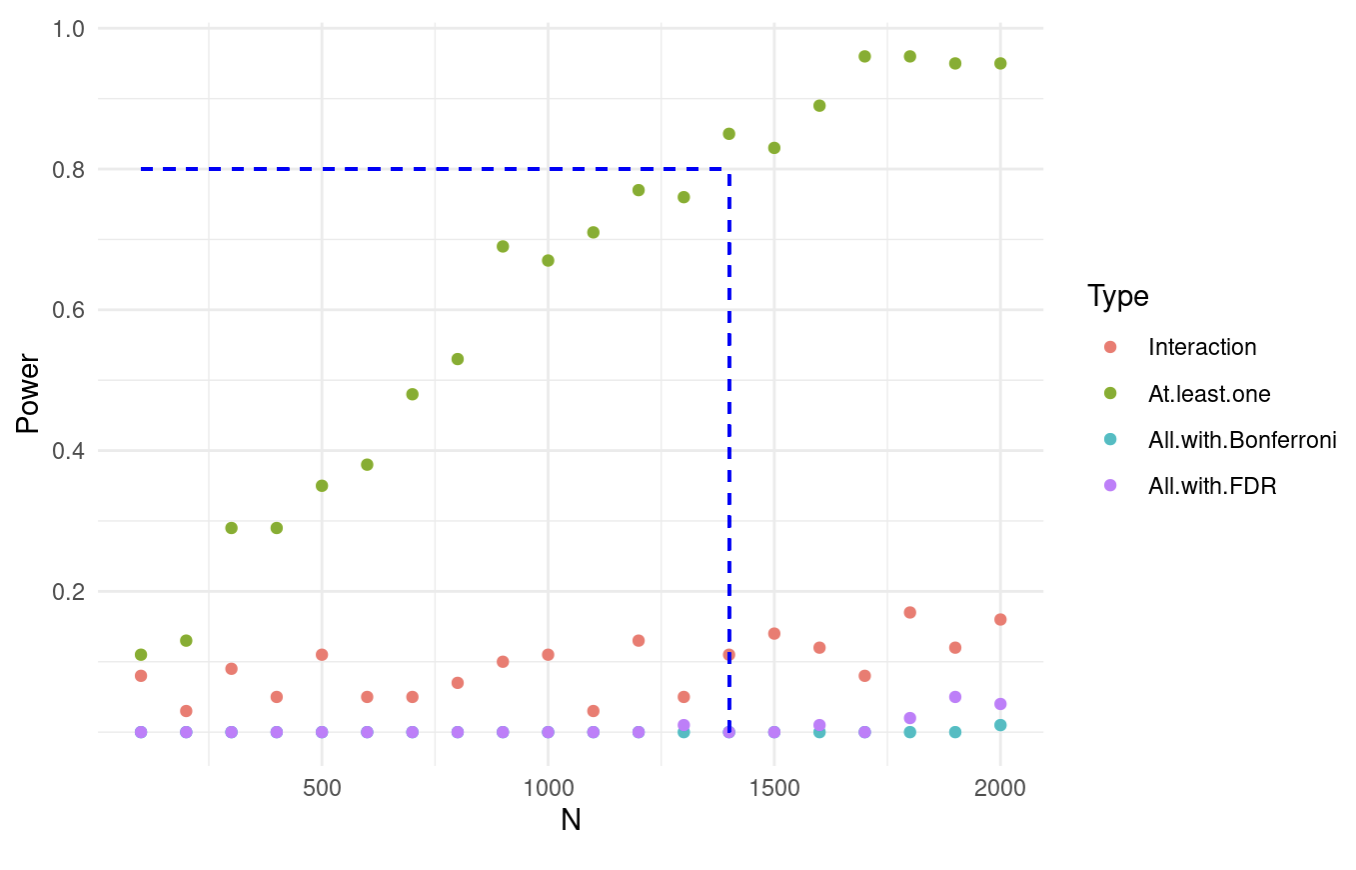
[PLEASE EXPAND ON RANDOMIZATION STRATEGIES IF NECESSARY, E.G., STRATIFIED RANDOMIZATION]

*TODO: modify, finalize; include considerations and any necessary graphs from slides 8-10 (including the graph of slide 7), and any additional necessary explanations (e.g. power). Add in explanation about stratified randomization.*

*Feedback:*

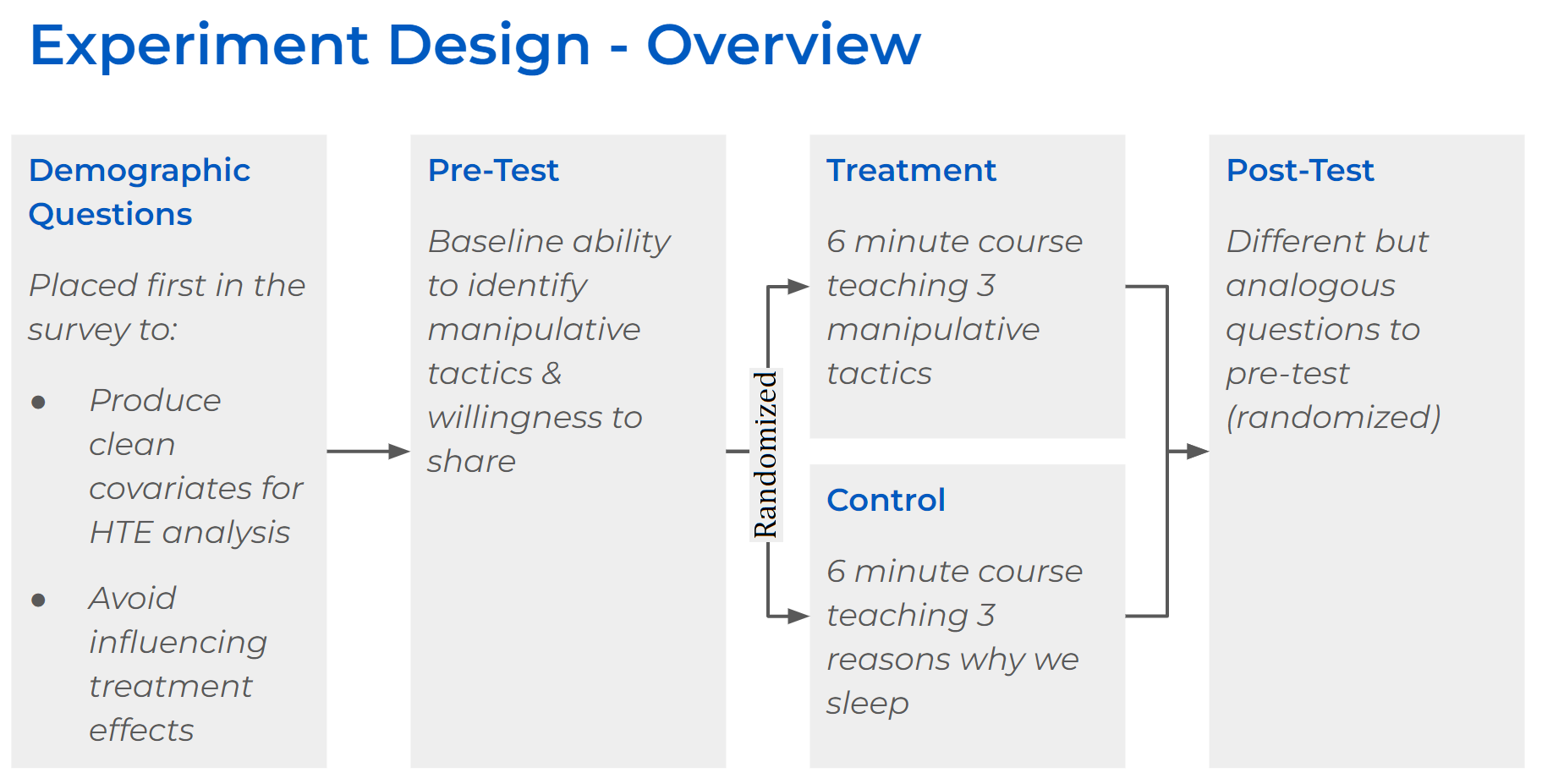
* *“Flowchart and graphics to describe experiment were helpful.”*
* *“You could benefit by explaining the experimental process more explicitly? Who is the sample? How would it be implemented?What does the treatment look like? What do the different posts look like that are used to measure outcomes? Please include them in the PAP.”*

Our experiment design focused on obtaining sufficient power to detect our main outcomes (i.e. identifying manipulative tactics), while also accommodating several HTE analyses on priority subgroups.



We performed a simulation using the pilot experiment design, testing for a statistically significant change in the confidence weighted accuracy score and 3 HTEs (4 outcomes in total) using only 1 treatment arm (Tactics). As shown in the figure above, a power analysis for the outcomes of interest suggests that 1,500 participants would give us less than 20% power for each hypothesis. (Since we are interested in being able to give precise recommendations, the “at least one”, the only powered approach is not appropriate for this context). This prompted us to modify our experiment to include a pre-post test design.

The following figure gives a high level overview of our new experiment flow:



This figure is explained in detail below.

*Demographic Questions*

The participants will first answer a set of questions to collect demographic information about age, gender, race, education, income, partisanship/ideology, trust levels, information sources, and social media use. (Please see the **Covariates** section and the **Survey Script** for more detail.) We placed the demographics questions before the pre-test since we care about using these covariates for HTE analysis. This way, the pre-test questions can not impact the participant’s reported demographics.

The demographic questions may prime participants to behave slightly differently during the pre-test (eg. they may be more aware that this survey tests misinformation since we ask questions about social media use). However, this effect should carry over to the post-test as well. Since we are measuring the difference between pre and post, and across the treatment and control groups, the priming effect should normalize out.

*Pre-Test*

After demographics, the participants respond to a pre-test that is representative of the tactics we will be teaching in the treatment arm. This sets a baseline for each participant to measure the incremental impact of our treatments. We also use this score as one of the covariates of interest for HTE analysis: susceptibility to misinformation at baseline. (See the **Pre-Post Test with Control** section for more detail.)

*Treatment/Control assignment*

After the pre-test, the participants will be randomly assigned to control and treatment groups. During the assignment, we will stratify based on income and ideology. That is, we will use Qualtrics to balance the random assignment of participants into each of the 30 possible income and ideology answer combinations (five options for ideology x six options for income). As a result, we can guarantee random assignment to treatment and control while also ensuring balanced assignment based on these two priority covariates. This procedure will enable us to further reduce variance and increase power. The ratio of assignment into treatment and control, within each stratum, will be a ⅔ treatment to ⅓ control. This uneven assignment should further increase our power since we expect the variance of the outcome to be smaller for the control (where the pre and post-test differences should be zero) than the treatment. See the **Power Calculation** section for more details on how we set this ratio.

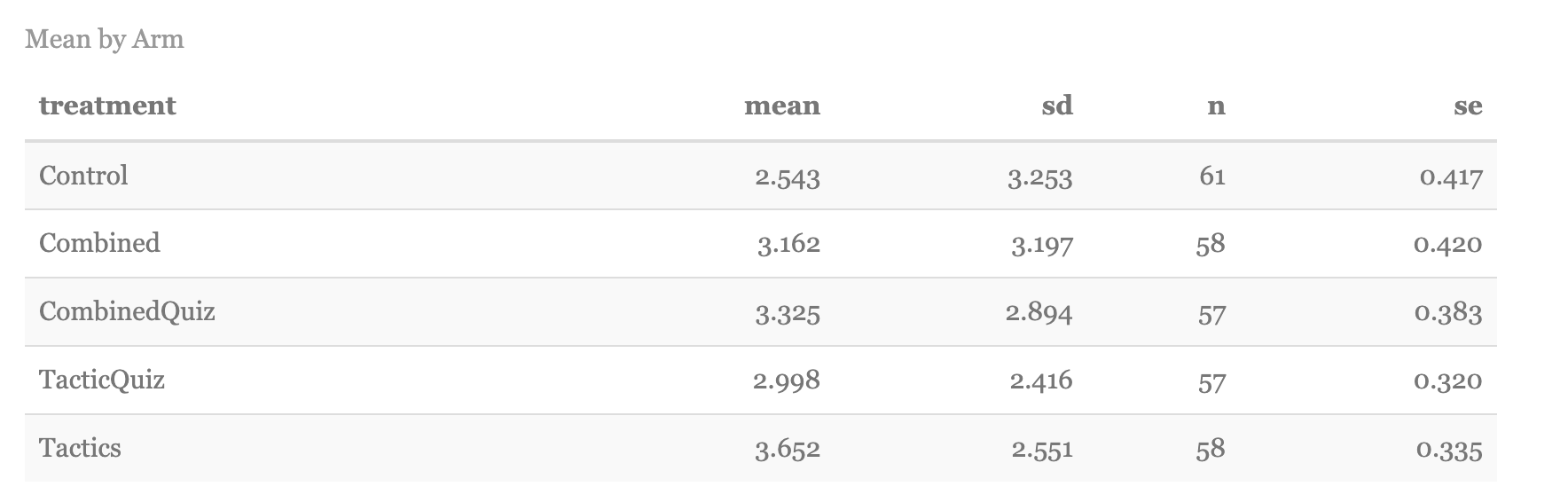
*Treatment & Control*

Both the treatment and the control intervention will contain a ~6-minute mini-course. For the treatment, the content will cover 3 manipulative tactics common to misinformation. The control group will see a mini-course on the neutral topic of sleep. (Please see the **Treatment** section for more detail.)

*Post-Test + Outcome Questions*

Finally, the participants will work through a post-test that has comparable examples as the pre-test. This will assess if the participants learned the tactics taught in the mini-course. (Note: we will randomize for each participant which question of the comparable pairs they answer pre and post to ensure that slight differences in the exact questions aren’t impacting the effects we identify.) The control group helps normalize the post-test improvements that come naturally from doing more questions. (Please see the **Outcome** section for more details.)

### **Pre-Post Test with Control**



Based on the figure above, the pilot suggested that the ability to recognize misinformation is highly variable between different people. The standard deviation (SD) of misinformation identification overall is ~3 units on a 20 point scale. This means a 1 SD confidence interval would span 30% of the scale of possible values. In order to maximize the power of our experiment and conduct HTE analyses, we decided to structure our experiment with a pre-post test along with a control group.

#### *Advantages: Increased Power*

The main benefit of conducting a pre-post test experiment is the increased power from (i) decreased variability and (ii) the greater proportion of treatment assignment relative to control.

With a pre-post test, we match the results of each individual with themselves and evaluate the change in their individual performance. This reduces the inter-person variability since the differences between each participant is accounted for when comparing to themselves. Decreased variance leads to higher power.

We also expect a significant decrease in variability in the control group. By designing the pre-post tests to be comparable to each other, we would expect individuals in the control group to perform almost identically on each of the tests. This low variability in the control group means we can shift a proportion of participants from control to treatment. More treatment units will permit HTE analyses of more subgroups. We estimated this drop in variability by running simulations on our pilot data and chose a conservative 1:2 split between control and treatment. (See **Power Calculations** for more details about this calculation and the reduction in variability.)

#### *Considerations*

There are 2 downsides to using a pre-post test design that we needed to account for.

First, it increases the length of the survey. Since we have 2 tests, the time taken to complete the evaluation is double the original pilot design. This means that we needed to cut down the number of questions we ask in each test to stay within our survey time limit. We chose to cut each test down to 6 questions.

Second, the pre and post-tests need to be comparable. Many misinformation pre-post experiments in the literature asked identical questions in each test (eg. Basol et al 2021). We worry that this would inflate the treatment effect since users would be answering the same question twice, and have the opportunity to reflect on their answers as they take the mini-course.

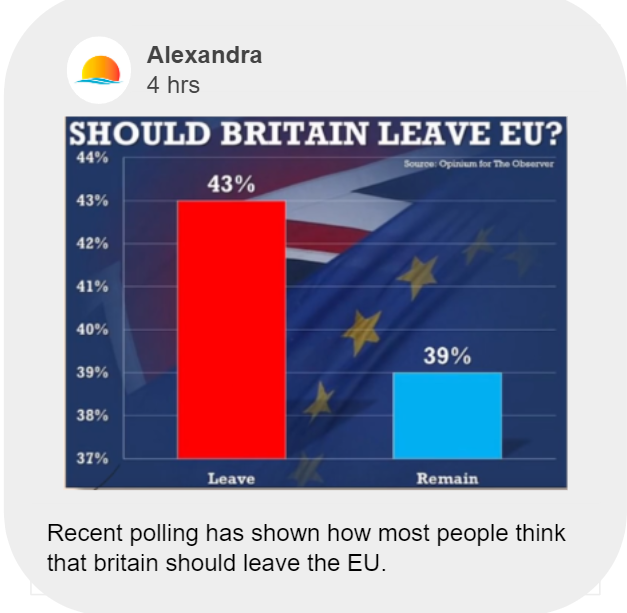
To ensure that we test the ability to recognize the tactics themselves, we design our pre and post-tests to contain different questions, presented in comparable formats using the same distribution of manipulative tactics. We did this by developing the 6 pairs of questions in the following ratios: 2 true, 3 tactics, and 1 “combined tactics” . We then developed 2 comparable question examples to form each pair and will randomize the allocation of questions between pre and post-test among the participants.

The “combined tactics” questions combine the anecdote and false comparison tactics. It also contains a non-misleading graph. This forces participants to consider multiple tactics for each question. It also means participants can not assume that manipulative text is always accompanied by a manipulative graph (and vice versa), when we ask them to identify which tactics are being used.

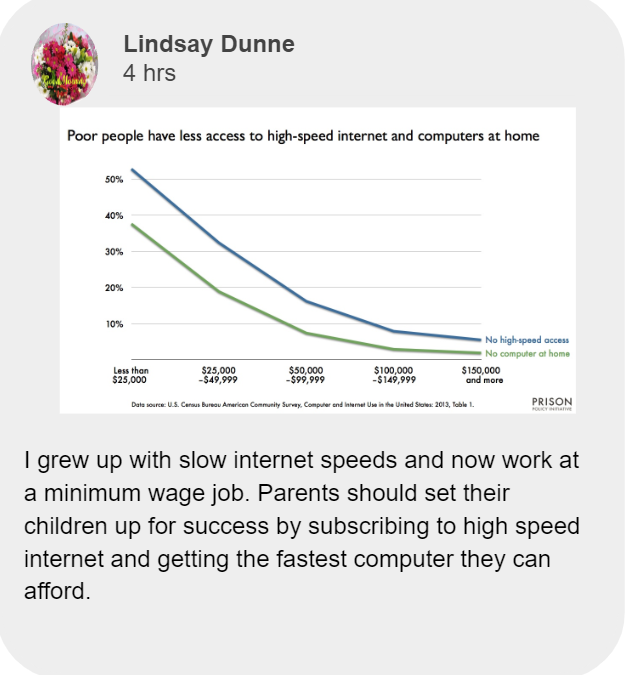
#### *Examples from our Pre-Post Tests*

Here are some examples of posts we show in our pre-post tests. For the full list, see our **Survey Script**.

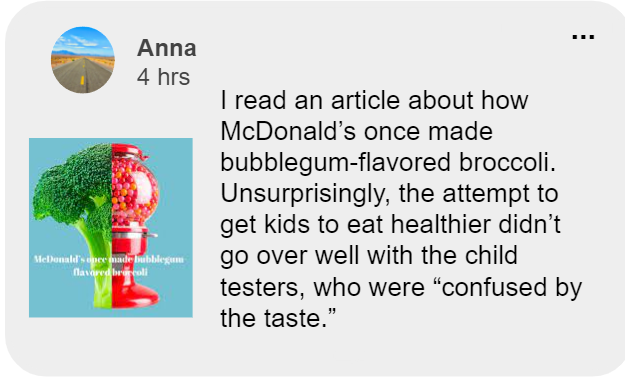
Misleading Graph:



Combined tactics:



Truth



### **Treatment**

*Please explain your experimental interventions in detail here. Please also provide rationale for choosing the treatments.*

*TODO: modify, finalize; considerations and any necessary graphs from slides 11-14 (including the examples from slide 13 and possibly image from slide 12), and explanations/rationale (e.g. effectiveness and sources - literature, pilot, partner)*

*Feedback: “Good motivation of treatments and the pre vs. post design on the qualitative side, but you could benefit from more quantitative analysis here.”*

In designing our treatment, our motivation was twofold. First, we wanted to maximize the effectiveness of our intervention, given the limited survey length. Second, we wanted to ensure that our experiment provides novel and actionable insights relevant to First Draft.

For reasons discussed in the **Experimental Design** and **Power Calculation** sections, we only have one treatment arm in addition to the control. Based on our discussion with First Draft, we decided that focusing on multiple outcomes and studying potential heterogeneities (see the **Covariates** and **Hypotheses** section) are the best use of power instead of testing various treatment arms. The outcomes we focus on include multiple behaviors in addition to the ability to identify manipulative content in general as well as specific manipulative tactics. Additionally, as discussed in the **Experimental Design** section, unlike the pilot, we include pre- and post-assessments of our outcomes of interest to maximize power.

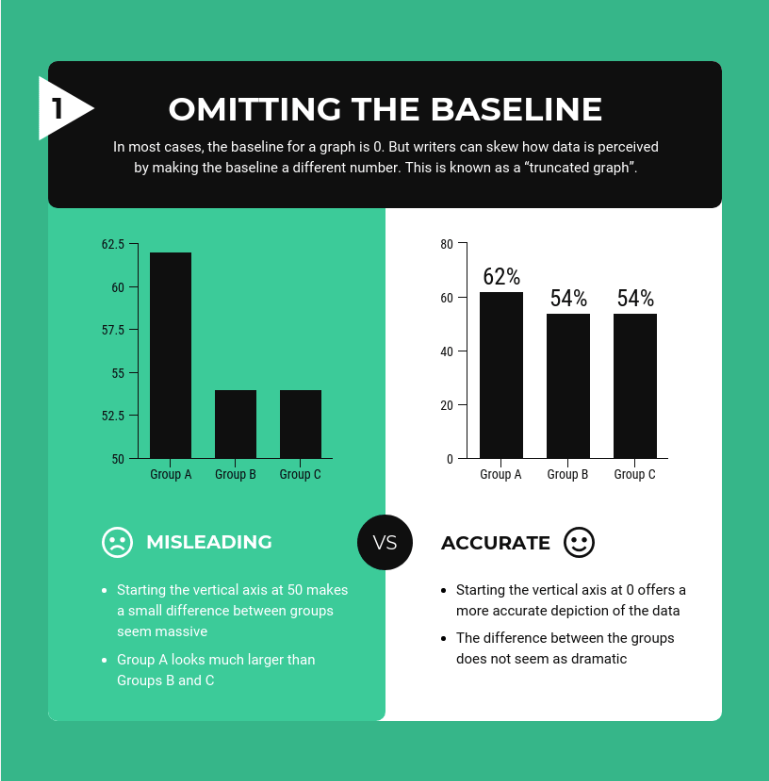
As discussed in the **Motivation** section, we will only have ~6 minutes to deliver the treatment. We therefore designed our treatment so we can still detect effects despite the course being shorter than First Draft’s SMS courses and the pilot.

First, we maximized participant engagement in the delivery of the course. Both First Draft and the literature suggested that interactive course elements increase participant engagement (Roblyer & Wiencke (2003), Hake (1998)). Interestingly, our analysis of the pilot data suggests that the pilot courses with quizzes didn’t outperform those without (there is also large uncertainty as the confidence intervals are greatly overlapping), see the figure at the end of this section. However, based on our literature review and conversations with First Draft, these quizzes only encouraged limited engagement – adding just a single quiz question at the very end of the treatment message. Therefore, following the literature, we included the following elements in delivering our message (Klein, Junior & Barbosa (2015), Ziden & Rahman (2013)):

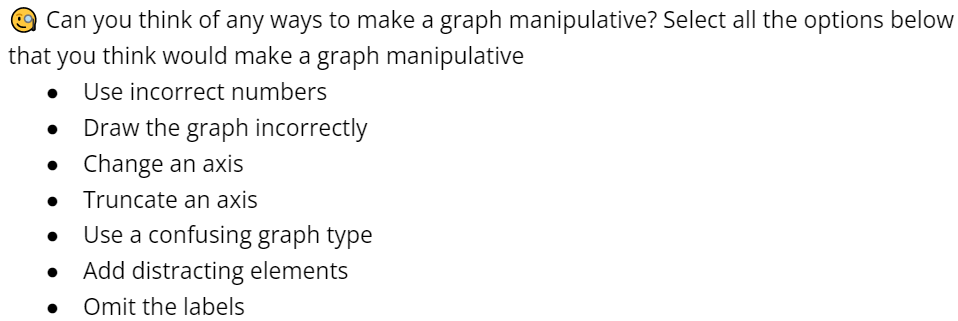
* Short and engaging, SMS style message snippets:



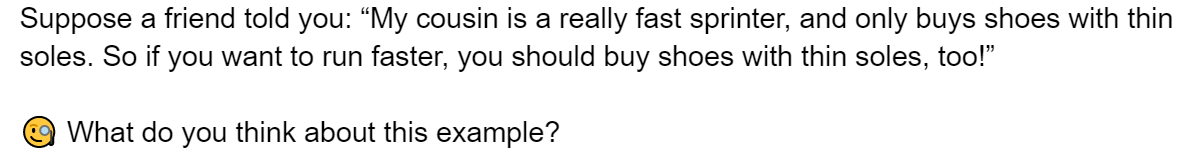
* Infographics (McCready, 2020):



* Interactive questions to provoke thought:

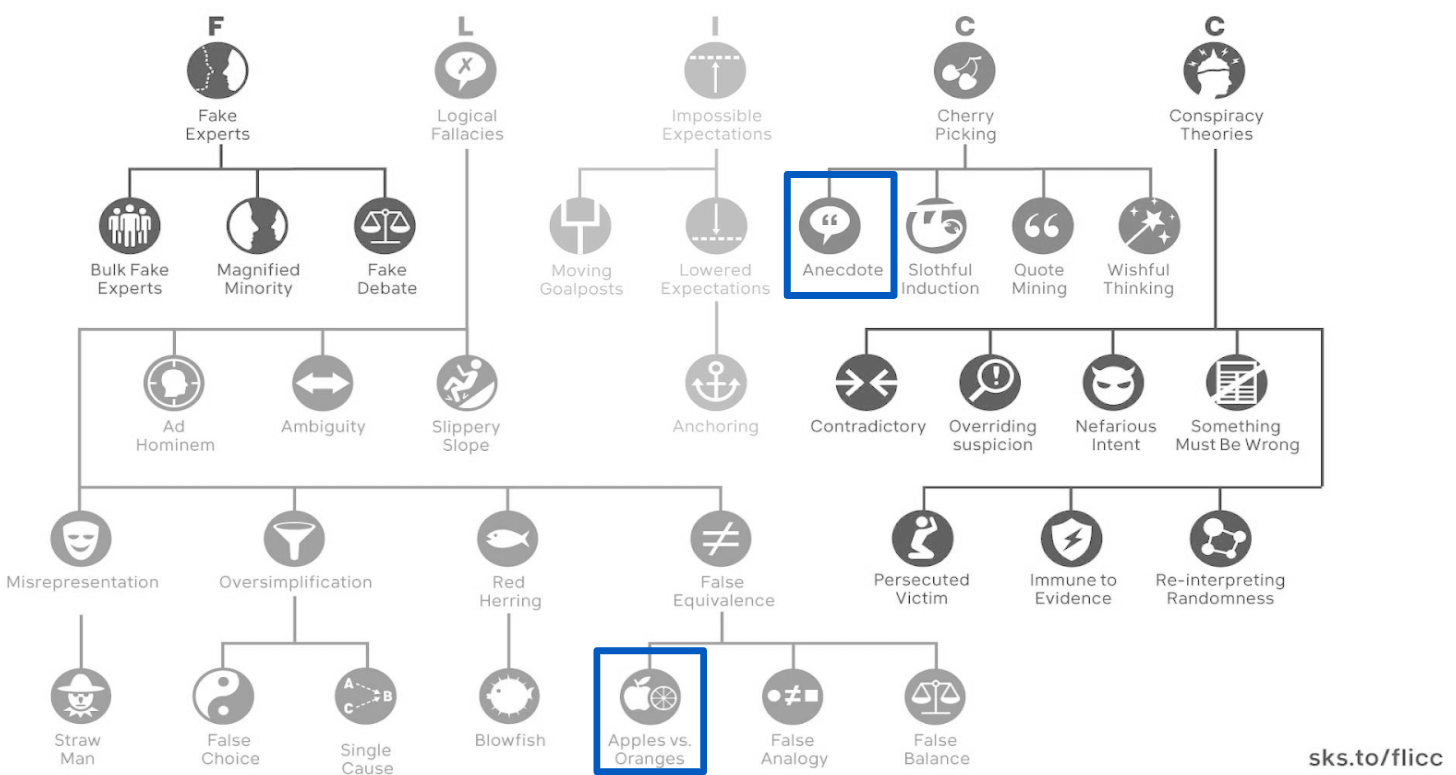


* Interactive and relatable example:

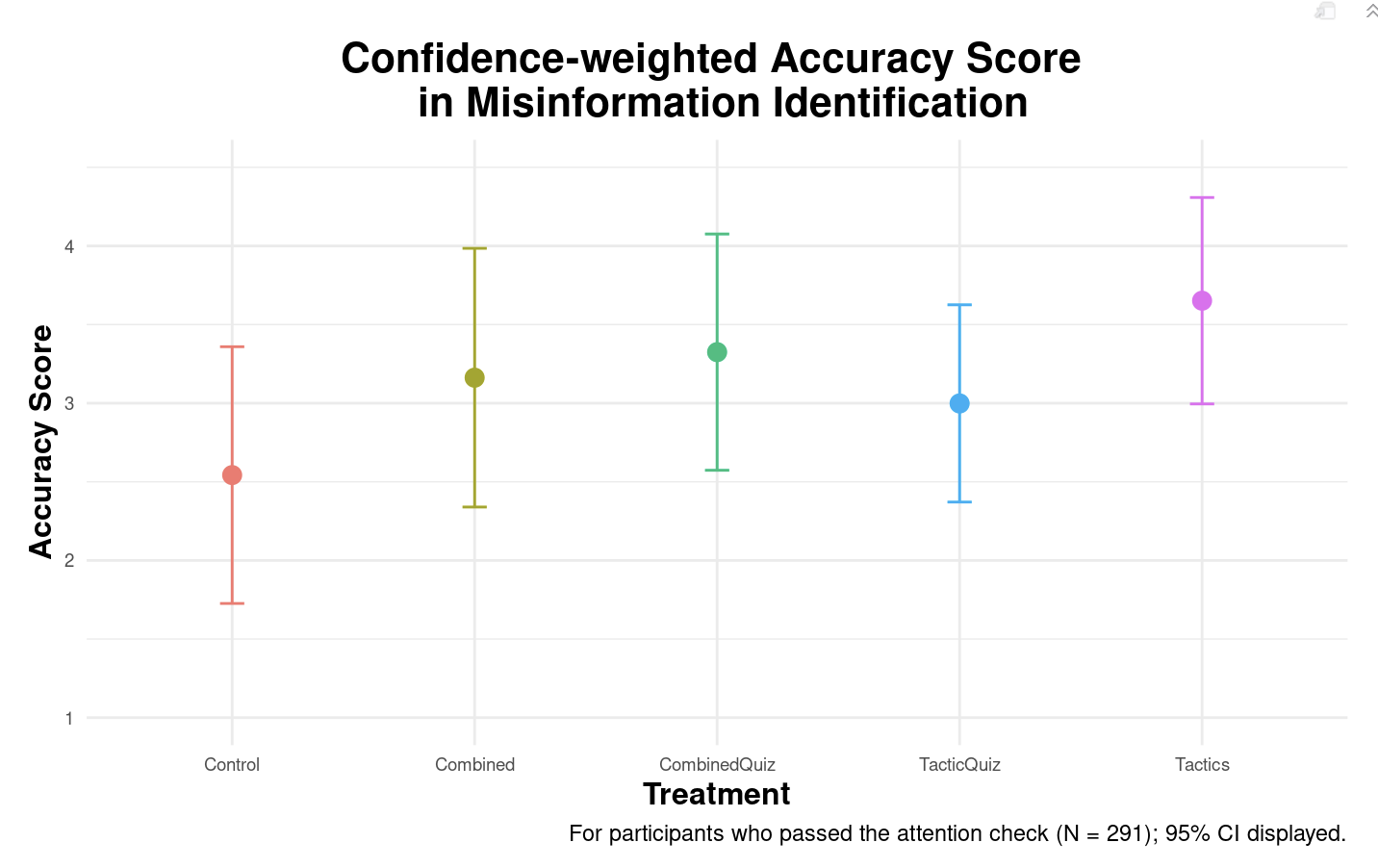


Second, within the broader context of misinformation inoculation, our course focuses specifically on manipulative tactics. We did this for several reasons. The tactics course seemed more promising than the emotions one and was included in all the pilot treatments either by itself or combined with the emotions course. As the graph at the end of this section shows, the pilot data suggests that the tactics course itself has comparable effectiveness (or better) than the combined one so it is a promising area of focus (although the confidence intervals are greatly overlapping). Additionally, based on our literature review, interventions focused on specific manipulative tactics have been very effective (Roozenbeek & Van der Linden (2019), Cook, Lewandowsky & Ecker (2017)). Furthermore, First Draft is very interested in learning about potential manipulative tactics that can be taught effectively, particularly tactics that are not yet widely studied. Finally, as suggested by our Partner, focusing on tactics rather than misinformation helps steer our course away from partisan topics which may alienate some participants. We expect the use of non-polarising language to maximize engagement and effectiveness in delivery.

We decided to focus on only three specific manipulative tactics because we have very limited time to deliver them. We picked one tactic already included in the First Draft SMS course and the pilot: misleading graphs. We chose two other tactics that are less studied in the literature and not yet in the First Draft course, because of the Partner’s above-mentioned request. These two were picked from the Cook (2020)’s FLICC taxonomy (shown below). We selected anecdotes and false comparisons because they: (i) are not dependent on context or prior knowledge, (ii) may be applied to a post on any topic, (iii) are not widely analyzed in the literature, and (iv) can each be feasibly taught in ~2-minutes.



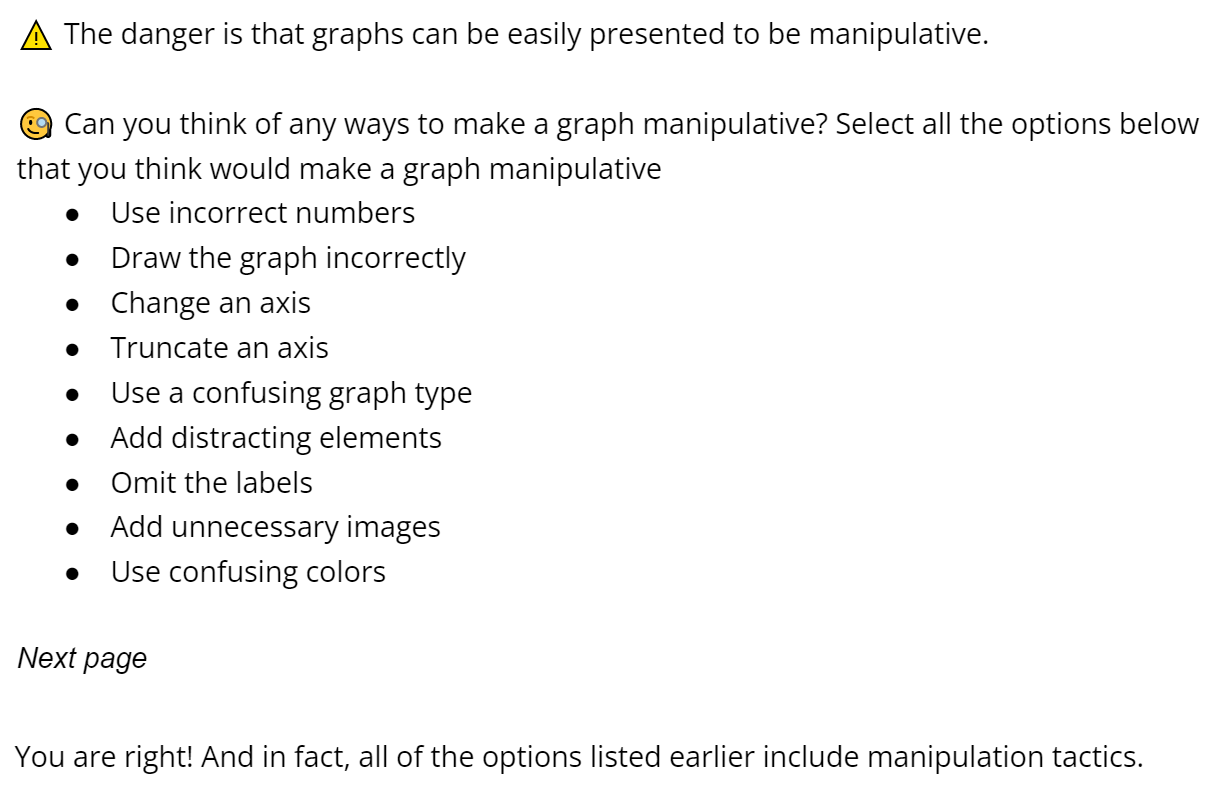
Our final consideration in treatment design is choosing a control message that is as neutral as possible. Our control messages use the same delivery style and interactive elements as the treatment. They are also educational but focused on a neutral topic unrelated to manipulative tactics and news, and that is also not too boring nor too exciting. We considered various options based on the literature (e.g. freezer burn in Lewandowsky & Van Der Linden (2021)). The best topic we found that fits these criteria is “sleep.” We are using specific pieces of information based on the book *Why We Sleep* by Matthew Walker.

  
*Summary of our analysis of treatment effects in the pilot*

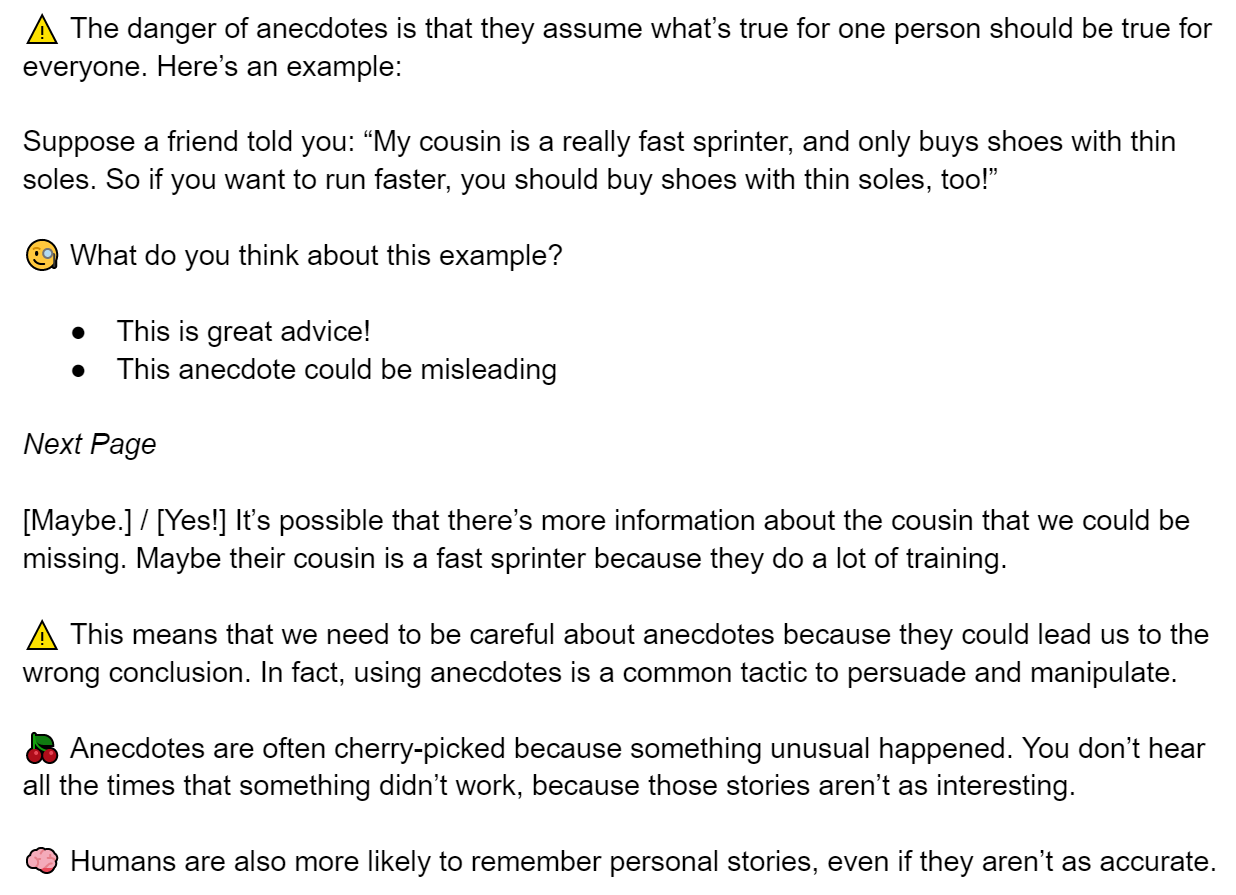
*You should provide details for each treatment arm here. Remember, visuals help! You can provide images or screenshots of what the actual experiment will look like to participants.*

We have included above the examples of what makes our delivery engaging and interactive. Below are some of the key pieces of information that we are teaching the participants about each of the manipulative tactics. For all the details of our treatment and control messaging, please see the **Survey Script**.

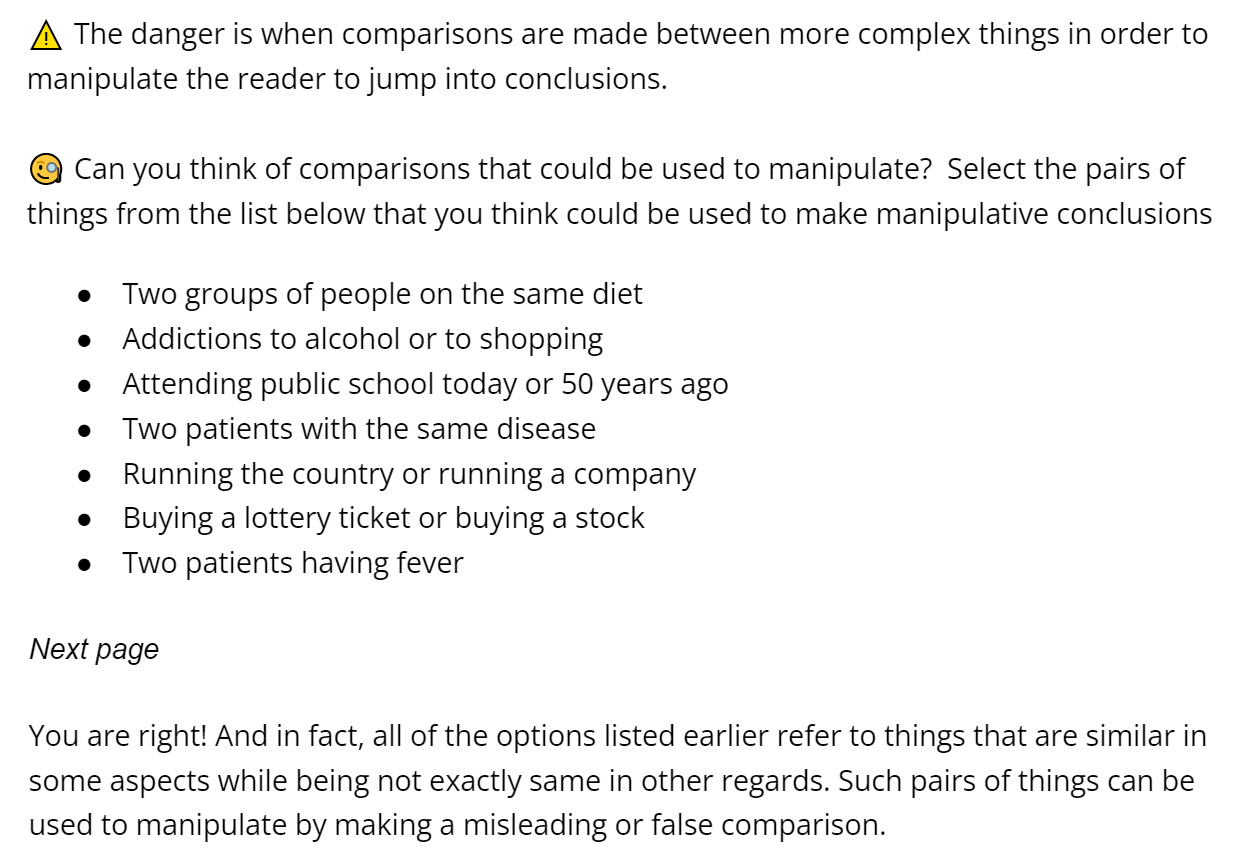
Key pieces of the misleading graphs content:



Key messages of the anecdotes content:



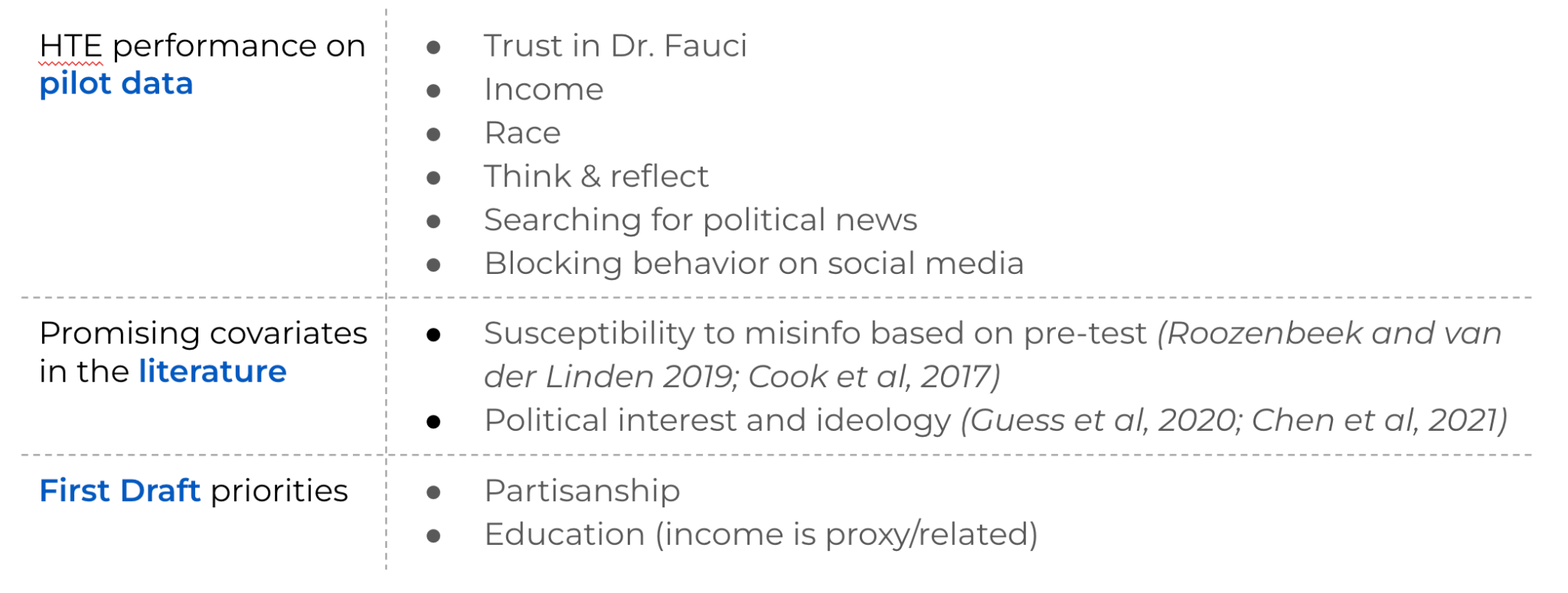
Key pieces of the false comparisons message:



### **Covariates**

*Please explain what covariates you will collect and use in your analysis.*

The pilot included a rich selection of covariates. The primary use of the covariates is to search for heterogeneous treatment effects in the population and to use them as control variables to reduce variance and increase power (to achieve this we also stratify the random assignment of participants into treatment and control to maintain covariate balance). Therefore, to decide which covariates to collect, we used the pilot data to identify important covariates to use while removing variables that are highly correlated with others. We also used the pilot data to identify particularly promising covariates for HTE analysis. We have supplemented the pilot's high-performing covariates with a literature review to identify additional covariates that show significant HTE in similar experiments. We also consulted our Partner about covariates deemed important. For identifying the covariates for HTE, the most promising ones from the pilot, the literature, and from First Draft are summarized in the figure below.

**

The following table shows a quantitative analysis of the covariates selected from the pilot selected from Tactics and Tactics Quiz treatment arms. All the listed covariates are significant at the 10% level. The results, however, cannot be trusted completely given the high variability and low power. Note, the Race covariate highlighted above was picked almost everytime by the causal tree package, which makes it a valuable covariate in HTE, even though it did not meet the criteria for the below table.

**

The mean and standard error of the confidence weighted accuracy score of the control population is 2.543 and 0.471 respectively. The estimates of HTE show substantial interactions with the treatments in the pilot. The inflation in interaction size can be attributed to the high variance in the study. Based on the above considerations, our primary focus for HTE are susceptibility to misinformation at baseline, income, and ideology. We chose susceptibility since it is found in the literature and the related variable, Think & Reflect, has been identified in the pilot as promising. We measure susceptibility to misinformation through the pretest score. Income stood out in the pilot and is related to education that is of interest to First Draft. Ideology is promising based on the literature, is of interest to First Draft, and is also related to the promising pilot variable, Trust in Fauci.

In addition to the above covariates considered for HTE, we collect the following set of demographic variables: age, gender, race, education, trust levels, information sources, and social media use. We will use these variables as control variables to reduce variance in our main regression analysis. We will also use these to identify HTE subgroups from data using causal trees. (Please see the Survey script and the Hypothesis section for more detail.)

### **Outcomes**

*Please explain what are your primary outcomes of interest, secondary outcomes of interest and why. If you collect multiple responses from the same individual, how would you select which ones to focus on or how would you synthesize them into a response function?*

*TODO: modify, finalize, include considerations based on slides 16 (include some of the considerations from the appendix, pilot, literature, partner), and any additional necessary explanations; also include how exactly outcomes are calculated (mentioned in slides 19 & 20)*

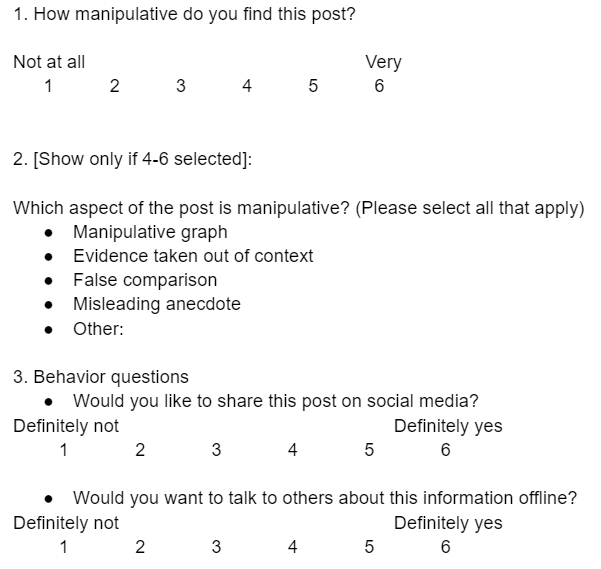
*Feedback:*

* *“Good choice of outcomes (tactics) as they rely less on substantive knowledge.”*
* *“Good discussion on outcome and covariate selections.”*
* *“Good thinking on outcome considerations, but again, we were looking for more quantitative evidence here. Again, please include them in the PAP.”*

Our experiment has two outcomes of interest:

1. Participants’ ability to identify manipulative tactics
2. Participants’ willingness to share manipulative content, both online and offline

We will measure these outcomes with the questions below, see the screenshot below. These questions will be asked for every post shown in our pre-post tests. (See **Pre-Post Test with Control** section for specific examples).



#### *Main Outcome - Detecting Manipulation*

We adjusted our primary outcome question to ask about identifying manipulative tactics rather than detecting misinformation since the latter often requires context and topic-specific knowledge. Manipulativeness is more objective and can be identified more consistently.

We also collapsed the primary outcome from 2 questions in the pilot to 1 in our new survey, with a 6 scale slider option. We chose 6 (as an even number) to force participants to choose whether they find a post manipulative or not. Simplifying this question saves time, but still measures participants’ confidence in their answer as in the original 2-part setup. This approach also is similar to many other experiments in the literature, including the Go Viral Game ([https://www.goviralgame.com](https://www.goviralgame.com/)) (Basol et al., 2021).

As mentioned in the Pre-Post Test with Control section, we calculated the estimated effect size from the results of our pilot. We saw a point estimate increase of over 20% between the treatment arms and control. Even with large SDs from the pilot, we can still reasonably expect at least 7-10% increase as our treatment effect, that we will be able to detect, since our treatment and our new outcome questions/tests are more targeted towards an objective problem. Also see more of this effect size discussion in our Power Calculations section.

#### *Tactics Identification*

We provide a multi-select tactics identification question that is only asked when the user selects a manipulative option for the main outcome (i.e. 4-6 on the 6-point scale). Through this process we will have the implicit selection for each manipulative tactic from each participant. We score this selection as correct or not to have our outcomes of interest.

#### *Behavior Questions*

We kept the pilot’s behavior question about sharing since sharing is one of main factors that accelerate the virality of misinformation. Regardless of users’ intentions, posts gain a far wider audience and experience network distribution effects when they are shared. As such, a reduction in sharing of misinformation is often the most effective way users can prevent its spread.

First Draft was also interested in measuring the influence of treatment on real-life behaviors. As such, we included a question about sharing information offline to capture actions that go beyond the social media platform.

We also considered many other behavioral questions, although they each suffered from several drawbacks. Examples include:

* **Would you fact check this content? Would you like to read more to find out the trustworthiness of this post?** In practice, fact-checking is a costly process for social media users. However, because answering this question is costless, participant responses may not actually reflect their true behavior.
* **Would you report this content to have it taken down?** Reporting behavior is relatively rare and many social media users aren’t aware it is even possible. Also, since the examples for our outcome questions cover deliberately neutral topics, we didn’t want to encourage users to report this type of content.
* **Would you unfollow users that post this type of content?** This action is a self remediation technique that could help users stop seeing posts from specific misinformation sources. However, there are many reasons to unfollow an account, and personal interest is one of the more common. Users may respond that they want to unfollow based on the topic of the examples we show them rather than their manipulative tactics.

### **Survey Script**

*Please include the entire survey script that will be deployed on Qualtrics.*

*TODO: finalize the script for qualtrics implementation, including the control text and finalized demographics*

The full text of our survey script can be accessed here: <https://docs.google.com/document/d/1t88qhy2M1ysVQHT0DTT5LrQTa2_ALSI1IWRADCECPn8/edit?usp=sharing>

## **Attrition**

We will have 3 attention checks: 1 during the demographics section, 1 at the end of the pre-test, and 1 before the post-test (after the treatment or control course).

We will include in our analysis all respondents for whom we have collected complete pre-test responses and have passed the two attention checks during the demographic and pre-test sections. As treatment is not revealed at this point, attrition should be independent of treatment assignment conditional on covariates.

For respondents who attrit after treatment assignment and before collection of post-test responses, the post-test answers will be coded as identical to the corresponding individual pre-test values. This procedure will enable us to identify unbiased intent to treat effects, although this procedure will attenuate the treatment effect estimates.

Additionally, we will test whether treatment assignment is predictive of attrition. If not, we will be able to estimate average treatment effects using the local average treatment effect on the treated, who did not attrit, without the above attenuation.

## **Hypotheses**

*Please lay out the specific hypotheses that you will test. Please also note whether they are the main hypotheses or secondary hypotheses. Please be clear about the measurement, and direction of hypothesized change. Please also provide rationale for making the hypotheses.*

*How do you plan to address the problem of multiple hypotheses testing? Do you plan to explore treatment effect heterogeneity and how do you plan to do that?*

*TODO: write this up nicely; include considerations and discussions based on slides 18-20, including the reference from the table based on the pilot analysis of covariates; reference previous secionts & include rationale for the hypotheses; specify which ones are secondary; HTE approach and explanations; correction for multiple hypotheses testing (include some of the considerations from the appendix, pilot, literature, partner)*

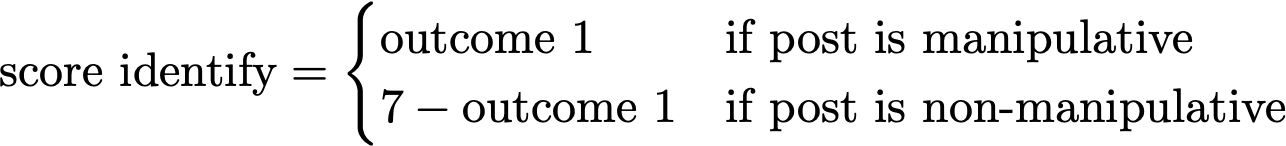
Our hypotheses are the direct reflection of our motivation, goal, and research questions (see the corresponding sections), as well as the considerations detailed in the **Experiment Design** Section. The first hypothesis is at the core of our project: ability to identify manipulative tactics. The second through fourth hypotheses focus on the ability to identify specific manipulative tactics, the three tactics covered in the treatment, to provide more granular insights into specific directions for future research and policy implementation. (Since our treatment combines messaging to target each of the three tactics, we cannot make causal claims about a particular messaging, e.g. anecdotes, being the most effective. We can only make causal claims about the course as a whole being the most effective for a specific tactic. That said, if the results for one of the tactics is significantly better, it can be taken as suggestive evidence or motivation for further exploring the corresponding part of the course).

The hypotheses five and six focus on the behavioral aspect of sharing misinformation. They are a direct translation of the considerations and motivation described in the **Outcomes Section**. Hypotheses seven, eight, and nine are modified versions of the first hypothesis; they are focused on estimating heterogeneous treatment effects on the ability to identify manipulative content. As described in the **Covariates Section** , our focus for the HTE analysis is susceptibility to misinformation at baseline, income, and ideology.

### **Main hypotheses**

**H1**: Participants will be more capable of identifying manipulative content after taking the course.

We measure the participants score using the following metric:



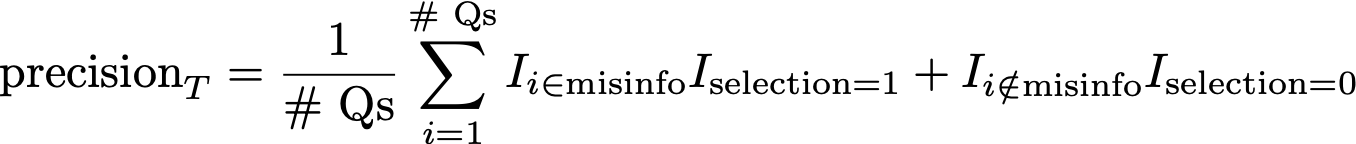
Our hypothesis translates into:



This is one of the main goals of the experiment and First Draft more broadly: making people better at identifying manipulative tactics related to misinformation.

**H2-4**: Participants will be more capable of identifying the 3 specific manipulation tactics covered in the course after taking the course.

We measure the ability to identify a tactic using precision



For each post, we code each tactic identified correctly (score of 1) if it is selected when that manipulative tactic is used or if it is not selected when the tactic is not used in the post. Otherwise, the tactic is not identified correctly (score of 0). Then, we take the average of the scores across the six posts for each tactic to find the precision score for each tactic.

This definition ensures that participants are penalized for not putting in effort, for example, by selecting all or none of the tactics for all questions.

Our hypothesis translates into:



We will test this hypothesis for each of our 3 tactics (graphs, anecdotes and false comparisons). This hypothesis satisfies the goal of measuring if the course taught participants how to identify individual tactics.

**H5**: Participants will be less likely to share misinformation online after taking the course

**H6**: Participants will be less likely to share misinformation offline after taking the course

We measure the likelihood of propensity to share misinformation by looking at the questions assessing manipulative content.



Our hypothesis translates into:



We follow the same strategy for the questions about both online and offline sharing behavior. These two hypotheses capture the propensity of misinformation spread both online and offline.

#### *Heterogeneity*

We expect our treatment to have varying effects on participants based on the respondents’ characteristics. To measure this variability, we will conduct tests for HTE. We limit our analysis to 3 HTEs in order to remain sufficiently powered. (See the **Covariates** section for our detailed rationale for focusing on the below three covariates.)

**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 will measure the susceptibility to misinformation from the pre-test scores.

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

We are asking the participants to pick political ideologies among 5 choices (very conservative, moderately conservative, moderative, moderately liberal, and very liberal) in the demographics section. Our discussions with the partner and the literature point to political ideology as an important covariate.

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

Participants will select their level of income in the demographic survey. Income showed high potential for HTE in the pilot.

### **Secondary hypothesis**

**SH1:** Participants will not identify true content as more manipulative after taking the course.

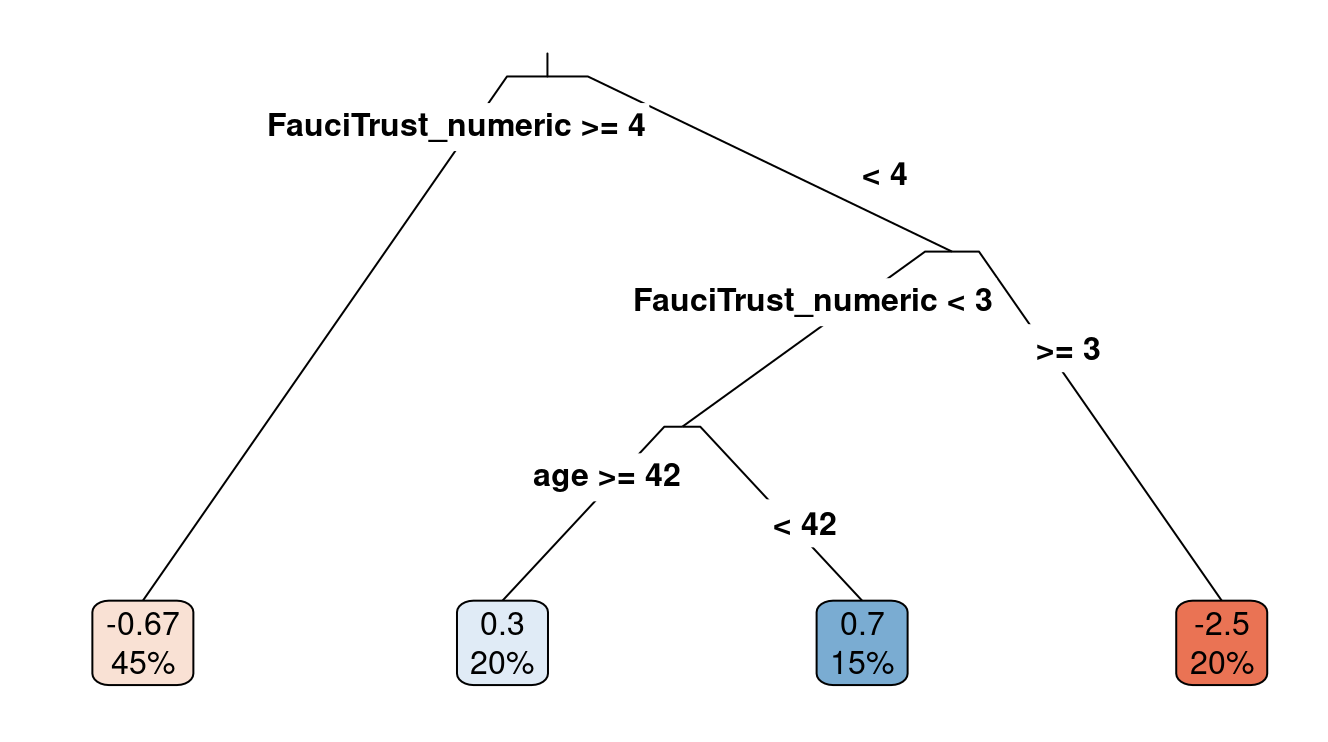
[ADD EQUATION IN RMD]

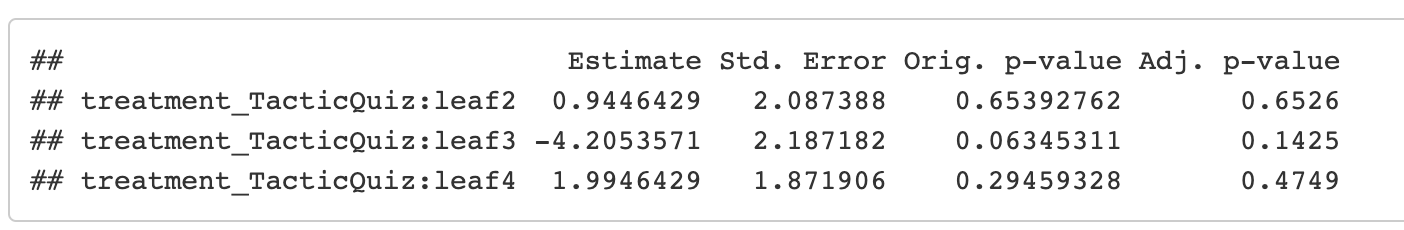
We want to ensure that our treatment is not making participants excessively distrusting of all content such that they start labeling true content as manipulative.

### **Exploratory hypotheses**

We will use the additional covariates from the demographic survey to form data-driven hypotheses using causal trees. These hypotheses will not inform the main conclusions of the experiment, and will act as additional findings to inform future research.

Our analysis on the pilot data for the Tactics and Tactics Quiz arms revealed some interesting findings that might be useful for future studies. Using multiple covariates to group people gives a better classification with very similar group members as compared to single covariate division. Therefore, we expect that groups made by causal trees will show stronger HTEs. This is evident from the pilot where we observed smaller *p*-values for HTEs on causal trees as compared to single covariate HTEs. The ML subgroups can also be used to inform new covariates or heterogenous policy assignments for future surveys. We also expect ML HTEs to reveal interesting correlations between covariates that can inform design decisions in the future.





### **Multiple Hypotheses Correction**

Since the hypotheses in our experiment are dependent on each other, a Bonferroni or Holm correction would be too strict. Benjamini-Hochberg correction is a better alternative, but it controls for False Discovery Rate (FDR) rather than Family Wise Error Rate (FWER) which is more substantive. We decided to use a bootstrap correction, by simulating our experiment from the pilot data with testing for what fraction of times we detected a false discovery. The code for bootstrap is given below. We found a corrected alpha of 0.0078 from our bootstrap simulation. This alpha performs FWER at a significance level greater than that of Bonferonni correction (0.05/9), which makes sense as the outcomes are dependent on each other. As the graphs show, we still have sufficient power.

## **Power Calculations**

*A crucial function of the PAP is to help you calculate the sample size you would need to detect a hypothesized treatment effect. Depending on your experimental design, please use the relevant tutorial for power calculation code.*

[*Here*](https://www.povertyactionlab.org/resource/conduct-power-calculations) *is a resource containing principles for running power calculations*

*Please indicate your choice of the minimum detectable effect and rationale. There is no universal rule of thumb for determining a “good” minimum detectable effect. For researchers, this might be informed by the existing literature: what have previous studies of comparable interventions found? What would be the smallest effect size that would be interesting to be able to reject? For partners, this might be the smallest effect that would still make it worthwhile to run this program (from their own perspective, or from a funder’s or policymaker’s perspective), as opposed to dedicating resources elsewhere. This may mean the smallest effect that meets their cost-benefit assessment, the smallest effect that is clinically relevant, or some other benchmark.*

*Please include the code, result, and figures from your power calculations.*

Below, we provide some code that you can use as examples to do your power analysis based on the design of your experiment.

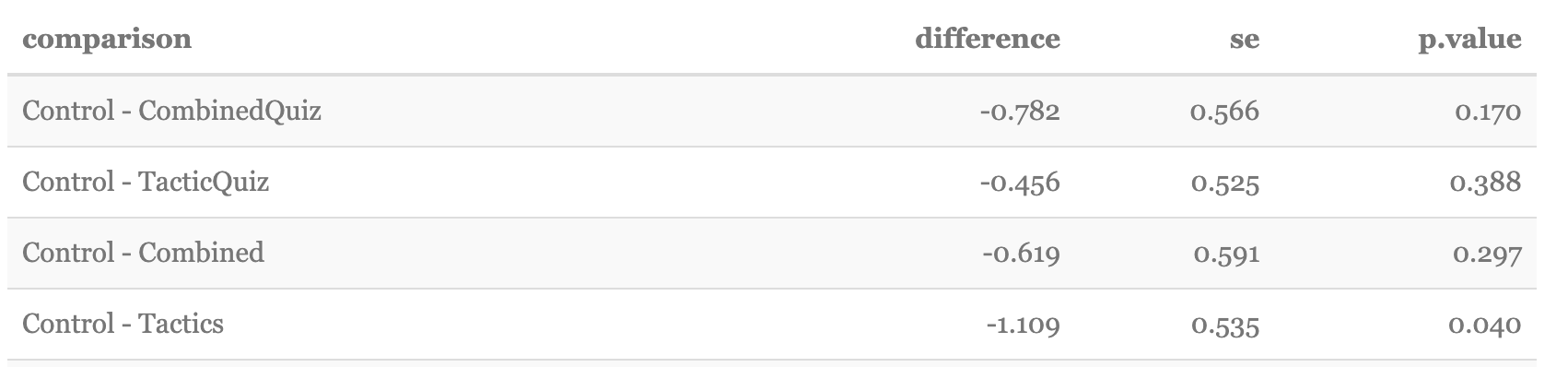
*TODO: write this up nicely; include considerations and discussions based on content and discussion of Slides 21-24; explain targeted minimum detectable effect and rationale (pilot, partner, literature), explain splitting the sample; explain the SEs used and how we got them, explain multiple hypotheses correction approaches and which ones we’re actually using - including considerations for correlated outcomes (including code, results and figures; include graphs/results for the different approaches we have considered); HTE approach (including reference to table with pilot results from above) and what the power calc results mean for our study*

*Feedback:*

* *“The heatmaps provided good visualization”*
* *“Good incorporation of advice from the TA (e.g., using control baseline from the pilot to do simulation as if it were a pre post design)”*
* *“The focus on power is good. You have done a lot of work here but it could be more clearly presented. For instance, you could have mentioned how you adjusted for sample size differences when you used pilot data to calculate power for your proposed design. Please include them in the PAP.”*

**Effect size estimation**

The pilot had an average effect size across all treatment arms of 0.74.



This is 30% of the baseline control score (= 2.54). We note that because of the relatively small sample size there is uncertainty around this number. Furthermore, the length of our mini-course will be shorter than the pilot. Papers too in the literature tend to find smaller effect sizes than the 30% from the pilot, for example studies with longer treatments (~1 hr) have shown ~20% effect sizes (Moore & Hancock (2022), Guess et al. (2020)). We expect a smaller effect size from a 6-minute course. Additionally, our discussion with the Partner revealed that a 10% effect size is more reasonable than 30%. Thus, to be conservative, we use the 10% effect size in our power calculations.

To conduct the power calculations, we took three main steps: 1. Variable estimation from the pilot study; 2. Simulations using these estimations; 3. Visualization of the results.

**Variable estimation**

In order to estimate variables for the power calculations (e.g. SDs), we transformed the pilot data to simulate our proposed test design. First, we rescaled the measurement of the outcomes to fall between 1 and 6 to match our proposed test. Next, the pilot study only measured post-treatment outcomes (not pre and post). To get an estimate for our test design, for the control group, we randomly labeled 5 (3 with misinformation, 2 true) of the 10 misinformation identification questions as being “pre-treatment” and the other 5 as being “post-treatment.” From there, we calculated the deltas between the means of the “post-treatment” and “pre-treatment” answers to calculate the expected mean and SD for the control group. These outcomes should be representative since the control group in our experiment will receive a 6 minute neutral treatment message that is designed to have no effect. In fact, these estimates are likely inflated compared to our actual experiment since our new pre-post tests are designed to be far more comparable. After these manipulations, the estimated SD for the control group based on the pilot data was 0.78 and the mean was ~0.

To get an estimate of the treatment group, we used nearest neighbor matching to match participants in the control group to individuals who received one of the treatments in the pilot study. Then, we calculated the average score for users in the control group as the “pre-treatment” stand-in and calculated the average score for users in the treatment group as the psuedo-post-treatment. Then, for the pairs of individuals, we took the delta between these averages and calculated a mean and SD that we used for the treatment group estimate for the power calculations. These measures give an approximation since the control group (who did not get any treatment) should be representative of individuals’ pre-treatment responses and the outcomes of the treatment groups in the pilot survey give us an estimate for the treatment group’s post-survey responses. Based on this method, the estimated SD for the treatment group was 0.99 and the mean was -0.35.

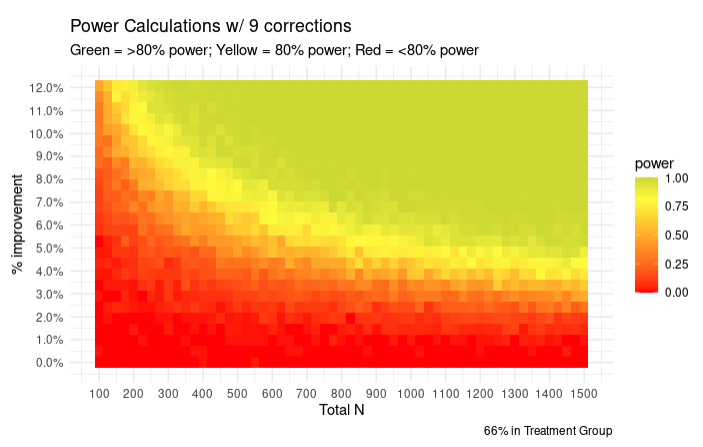
**Simulations**

We used the variables above as baselines for our power calculations. Given our treatment and control groups have different SDs, we used simulations for our power calculations. These simulations created synthetic data based on a given sample size, treatment-control split, treatment mean, treatment SD, control mean, control SD, and number of *p*-value adjustments (using bootstrap). Simulations were then run using these parameters to estimate the power of a given configuration of parameters.

**Visualization**

We used the power calculations to inform three core decisions.

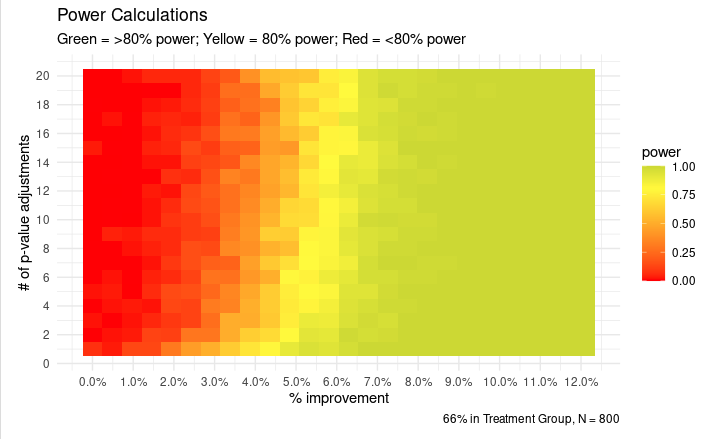
First, we examined the tradeoff between the total number of people in the experiment and the treatment effect. Visualizing these results suggest we can have a well-powered experiment with 800-1,000 participants and a ~7% improvement in the treatment groups.



Next, given we expect a lower SD in the control group compared to the test group (as discussed above, because of the design of the pre and post tests in our experiment, we expect no difference between the before and after scores for individuals in the control group), we used our power calculations to look for an optimal split between the treatment and control groups. Power analysis in the figure below suggests that a ⅓ vs ⅔ treatment-control split is plausible given a 7% improvement in treatment groups. Even though this split is slightly optimistic given our result below, we do think it is justifiable since we expect our control group to have lower SD than the pseudo calculations from the pilot as our pre and post questions are designed to be comparable. Moreover, following the guidance in Özler (2021) also leads to the ⅓ :⅔ split assuming a 0.99 SD for the control group and a 0.5 SD for the treatment group (the latter slightly smaller than our estimate of 0.78 above).



Finally, we examined the tradeoff between treatment effect and the number of *p*-value adjustments we could conduct. We settled on being able to adjust for 9 multiple hypothesis tests given our anticipated effect size.



## **Analysis**

*What analysis would you run? How to interpret the results?*

**Source: The Data Analysis and Experimental Design Tutorial contains code for analyzing the pilot data, including data cleaning, balance checks in treatment assignment, and analysis using regression models. The data collected from the actual experiment will have a similar structure, so please test the analysis out in line you propose over the pilot data as much as you can.**

[PLEASE INCLUDE YOUR ANALYSIS OUTLINE HERE IN WORDS.]

*TODO: write this up nicely; include considerations and discussions based on content and discussion of Slide 25 and reference the above sections as it makes sense*

[ADD CODE CHUNKS FOR THE STEPS]

We will start by cleaning the data and constructing the above mentioned variables for our outcomes and covariates. We will drop participants who fail the attention checks prior to treatment assignment. Additionally, we will perform balance checks on the covariates to check the random assignment.

Next, we will calculate summary statistics pertinent to our outcomes (formulas in the **Hypothesis** section). We will use regression analysis on the outcome metrics to test our central hypotheses (H1-H6). Additionally, we will check for HTE for the three covariates (income, political ideology, and misinformation susceptibility at baseline) as part of testing hypotheses 7-9. Note that we will be using the corrected value of significance level (0.0078) from our bootstrap simulations for our multiple hypotheses correction. As part of a different learning outcome of this course, we will also use the causal tree function to find data-driven hypotheses for additional HTEs.

Finally, we will do a power calculation of our main hypotheses through simulation to study the effectiveness of our experiment design.

## **References**

[PLEASE INCLUDE RELEVANT REFERENCES HERE IF APPLICABLE.]

*TODO: add in based on what is cited above*

* Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & Linden, S. van der. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. In *Big Data & Society*.
* Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. In *PloS one*, 12(5), e0175799.
* Cook, J. (2020). Deconstructing climate science denial. In *Research Handbook on Communicating Climate Change*. Edward Elgar Publishing.
* Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. In *Proceedings of the National Academy of Sciences*, 117(27), 15536-15545.
* Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American journal of Physics, 66(1), 64-74.
* Klein, A. Z., Junior, J. C. D. S. F., & Barbosa, J. L. V. (2015). M-learning in practice: Using SMS for teaching and learning in undergraduate courses. IEEE Latin America Transactions, 13(1), 321-329.
* Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology, 32(2), 348-384.
* Moore, R. C., & Hancock, J. T. (2022). A digital media literacy intervention for older adults improves resilience to fake news. In *Scientific reports*, 12(1), 1-9.
* Roblyer, M. D., & Wiencke, W. R. (2003). Design and use of a rubric to assess and encourage interactive qualities in distance courses. In *The American journal of distance education*, 17(2), 77-98.
* Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. In *Palgrave Communications*, 5(1), 1-10.
* Walker, M. (2017). Why we sleep: Unlocking the power of sleep and dreams. Simon and Schuster.
* Ziden, A. A., & Rahman, M. F. A. (2013). Using SMS quiz in teaching and learning. In *Campus-Wide Information Systems*.
* Özler, Berk. “When Should You Assign More Units to a Study Arm?” *World Bank Blogs*, 21 June 2021, https://blogs.worldbank.org/impactevaluations/when-should-you-assign-more-units-study-arm.
* McCready, Ryan. “5 Ways Writers Use Misleading Graphs To Manipulate You [INFOGRAPHIC].” *Venngage*, 17 Apr. 2020, https://venngage.com/blog/misleading-graphs/.