# # Introduction

Online misinformation has become a crucial threat to a well-informed citizenry. False and 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 focused on identifying a low-cost way of increasing people’s ability to recognize online misinformation. In particular, we implemented an intervention to teach people to recognize manipulative tactics in online information, and reduce their likelihood of sharing manipulative content online and offline.

We implemented a 15 minutes long online survey that included a 6 minutes long treatment as well as comparable pre- and post-intervention questions. The intervention messaging was designed based on First Draft’s SMS course about “[Protection from deception](https://firstdraftnews.org/sms-course/).” One-third of our treatment drew directly on topics covered by this course – specifically focusing on misleading graphs. The other two-thirds of our treatment explored new misleading tactics – anecdotes and false comparison – not previously taught in First Draft’s courses. In the design and delivery of the survey we focused on maximizing power for identifying an effect. For details on the design and our considerations, please see our pre analysis plan [https://drive.google.com/file/d/1ju8i62mImEzKa-X-o3MZ5x4EF7Fb03ge/view?usp=sharing]

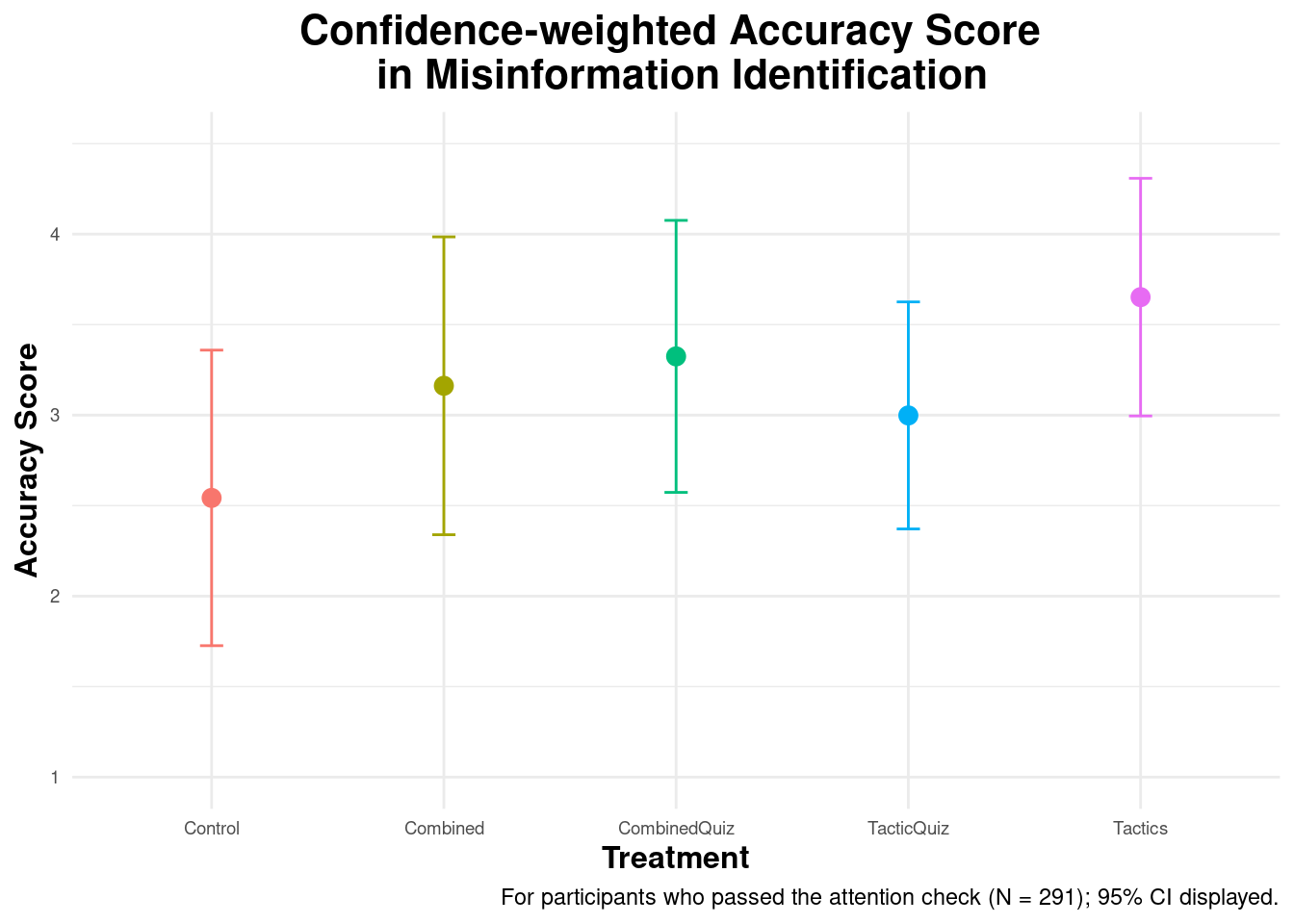
We found that our 6-minutes tactics focused course helped participants label manipulative posts as 12 percentage points more manipulative. At the same time, we saw increased skepticism for posts in general. Participants labeled non-manipulative posts as 15.4 percentage points more manipulative. In particular, participants were more skeptical of non-manipulative graphs. We see similar results with regards to ability to correctly identify the presence or absence of specific manipulative tactics. The 3 tactics taught were identified more after the treatment: there was a statistically not significant increase in false positive identification of these tactics as well as a statistically significant increase in identification of these tactics when the tactics are present for misleading graphs and anecdotes and a not statistically significant increase for false comparisons. We found no significant effects on sharing behavior and our results stand up to various robustness checks.

## ## Motivation

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. Our goal for this project was to study the efficacy of a low cost intervention like this, with potential to reduce susceptibility to misinformation and related sharing behavior online and offline. Furthermore, we aimed to explore content beyond the current First Draft course content, such that we could provide useful insights for First Draft to further improve its programs. To achieve these objectives, we reviewed the literature, the pilot results, and consulted with First Draft.

Within the broader context of misinformation inoculation, we decided to focus specifically on manipulative tactics. We did this for several reasons. Based on our literature review, interventions focused on specific manipulative tactics have been very effective (Roozenbeek & Van der Linden (2019), Cook, Lewandowsky & Ecker (2017)). Moreover, 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. Additionally, focusing on tactics rather than misinformation helps steer our course away from partisan topics which may alienate some participants. Finally, First Draft is very interested in learning about potential manipulative tactics that can be taught effectively, particularly tactics that are not yet widely studied.

Moreover, in the pilot, the tactics course seemed more promising than the emotions one. As the graph below shows, the pilot data suggests that the tactics course itself has comparable effectiveness (or better) than the combined one (which included tactics and emotions material) so it is a promising area of focus (although the confidence intervals are greatly overlapping).



## ## Research Questions

In light of our motivation and decision to focus on teaching participants about specific manipulative tactics, we decided to focus on the following research questions for our study:

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.

# # Methodology

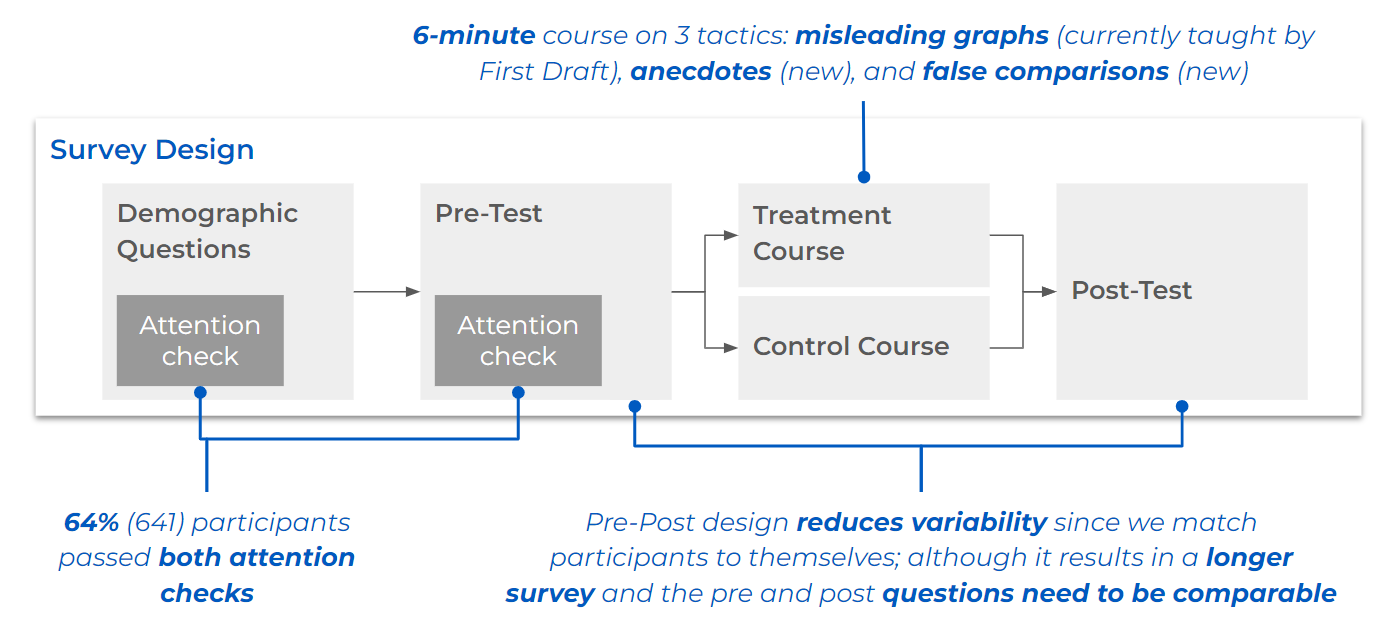
## ## Overview

Our experiment was 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 as 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 participants ). First Draft also indicated that they expect the treatment effect to be small.
* Taking these together, we were wary of the experiment’s power. (prior to attention checks). First Draft also indicated that they expect the treatment effect to be small.

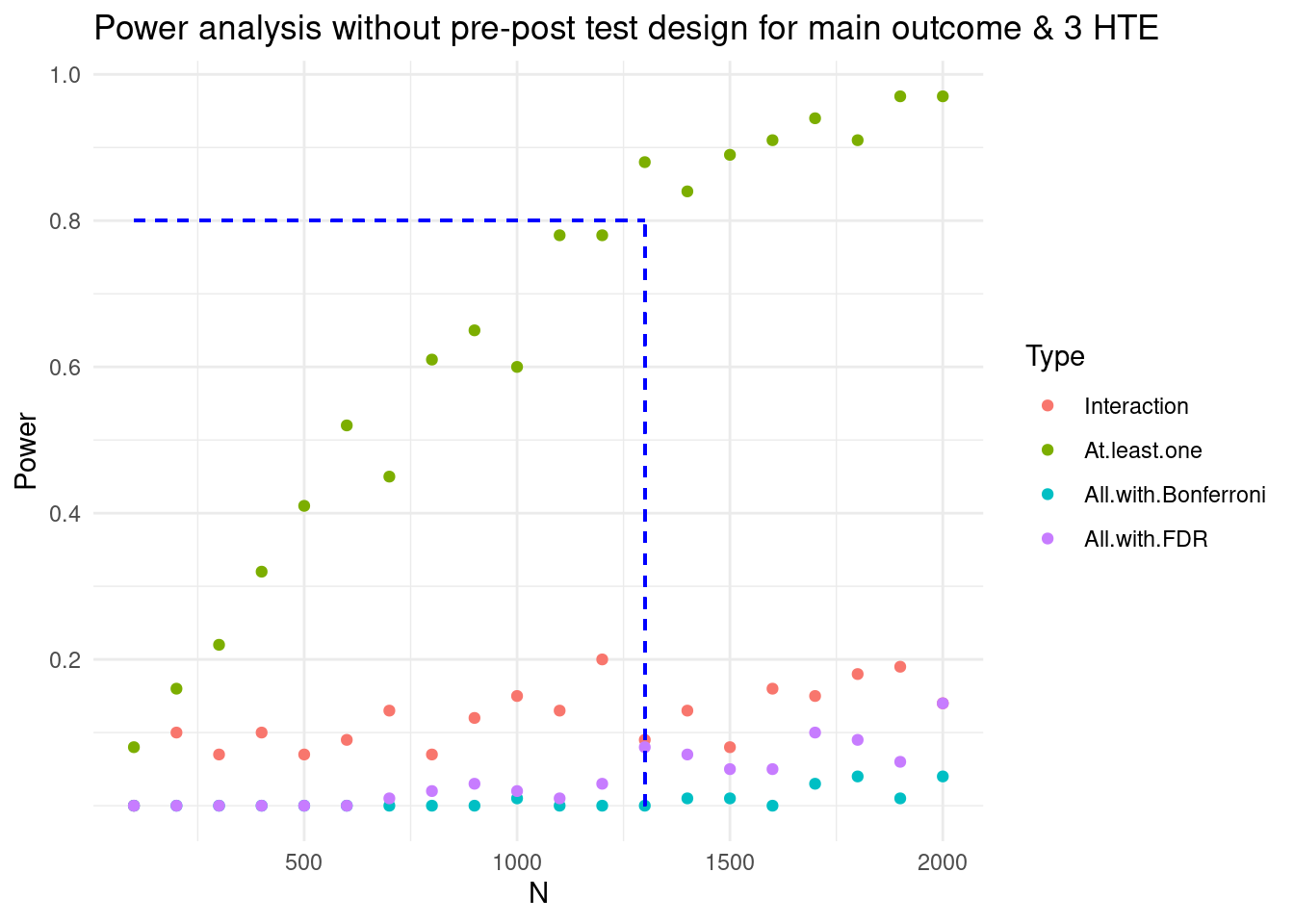
Taking these together, we were wary of the experiment’s power. We focused on obtaining sufficient power to detect our main outcomes (i.e. identifying manipulative tactics), while also accommodating several HTE analyses on priority subgroups. To do this, we estimated the experiment’s power by leveraging the results from the pilot study. See details about this in the following section and in our pre-analysis plan [https://drive.google.com/file/d/1ju8i62mImEzKa-X-o3MZ5x4EF7Fb03ge/view?usp=sharing].

See below the overview of the participants’ experience during the study. This is explained in more detail in the sections that follow.



## ## Pre and Post Design

[Insert code to generate graph]



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 giving precise recommendations, the “at least one” approach isn’t appropriate for this context.)

The above results prompted us to modify our experiment to include a pre-post test design. After asking the basic demographics questions, the participants respond to a pre-test that is representative of the tactics we will be teaching in the treatment arm. The answers to the pre-test questions form 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.

Both the pre- and the post-test involve 6 comparable social media post examples each. They see one example in each of the following categories:

* True (without a graph)
* True (with a non manipulative graph)
* Misinformation (with a manipulative graph)
* Misinformation (with manipulative anecdote)
* Misinformation (with false comparison)
* Misinformation (with manipulative anecdote, false comparison, and non manipulative graph).

We created two social media post examples for each of the six above categories and for each participant and each category we randomized which one they see in the pre and which in the post test. Additionally, we randomized the order of the posts for both the pre and the post test. To see the details of the social media post examples used, please see our survey script here: <https://docs.google.com/document/d/1t88qhy2M1ysVQHT0DTT5LrQTa2_ALSI1IWRADCECPn8/edit?usp=sharing>

Additionally, we decided to include two attention checks prior to treatment assignment to ensure quality answers and further reduce noise in our data.

## ## Outcomes of interest

Our outcomes of interest are based on the study’s four research questions. After each social media post example, we ask the same set of questions corresponding to our outcomes of interest. These questions are:

1. How manipulative do you find this post? *(6 point scale from “Not at all” to “Very”)*
2. Which aspect of the post is manipulative? *(multi-select)*
   1. Post not manipulative
   2. False comparison
   3. Misleading anecdote
   4. Manipulative graph
   5. Evidence taken out of context
   6. Other (require text entry)
3. Would you like to share this post on social media? *(6 point scale from “Definitely not” to “Definitely yes”)*
4. Would you want to talk to others about this information offline? *(6 point scale from “Definitely not” to “Definitely yes”)*

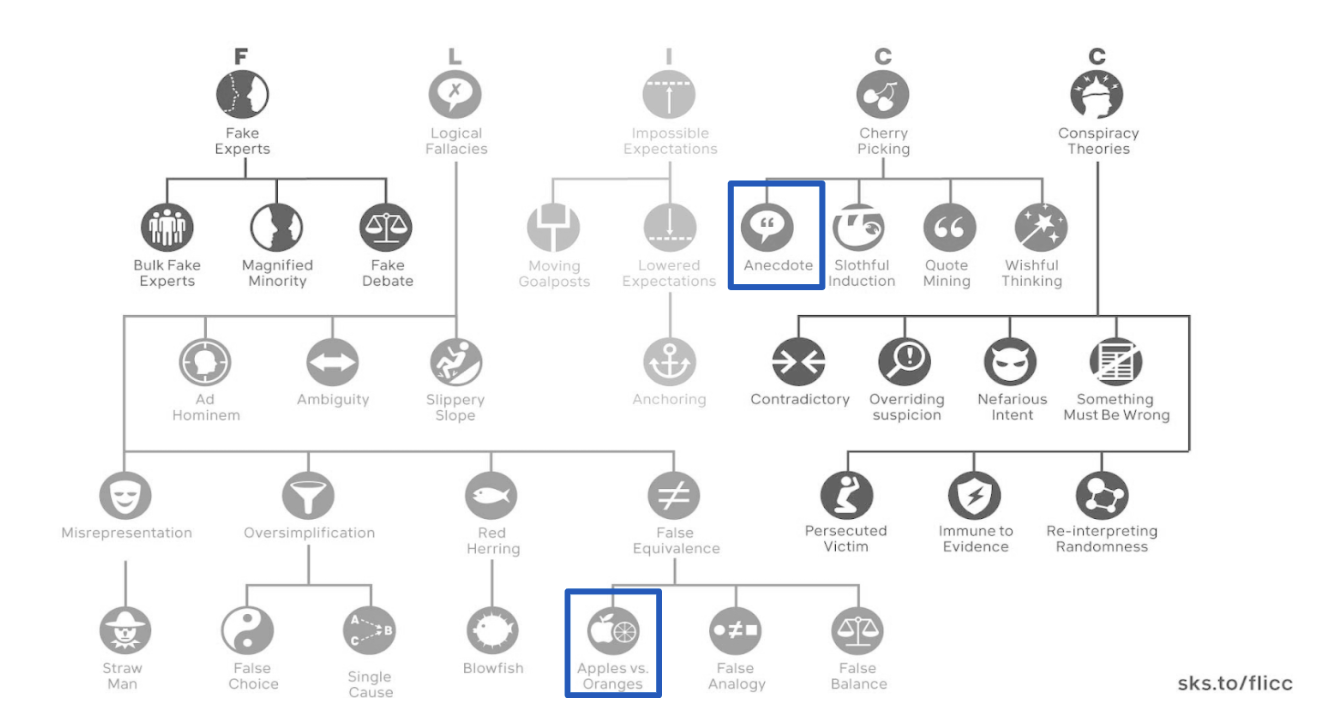
We use the answers to the first question across all six post examples to investigate H1, SH1, and the HTE related H7-9. (We look at the difference between the answers to the six posts before versus after the treatment or control messages).

For H2-4 about the ability to identify specific manipulative tactics, we use the selected answers for the second question, across all six post examples. (We look at the difference between the answers to the six posts before versus after the treatment or control messages). When evaluating whether or not a specific tactic was correctly identified for a given post, we compare whether the tactic was selected or not against whether the tactic is actually present or not. In general, for correct answers we expect the tactic to be selected when it is present, and unselected when it is not present. However, we explore multiple ways of scoring these answers that focus on subtle differences in what we care about: for example, always having exactly the correct answer or making sure that the tactic is selected when it is present. See the Metrics Definition section of the Tactics sections of Results for more details.

We use the answers to the third (fourth) question across the four manipulative post examples to investigate H5 (H6). We look at the difference between the answers to the four posts before versus after the treatment or control messages.

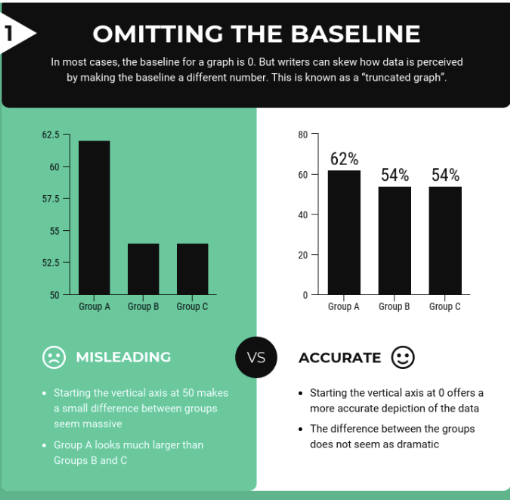
## ## Intervention

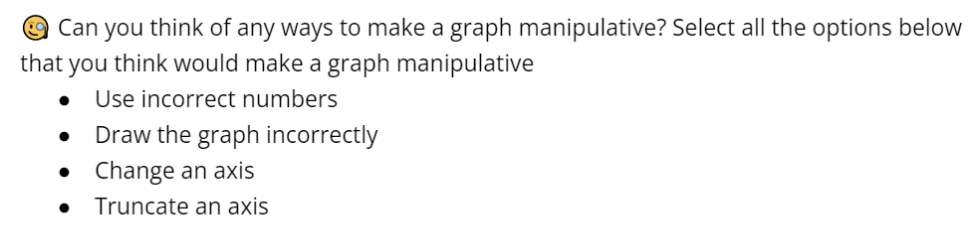
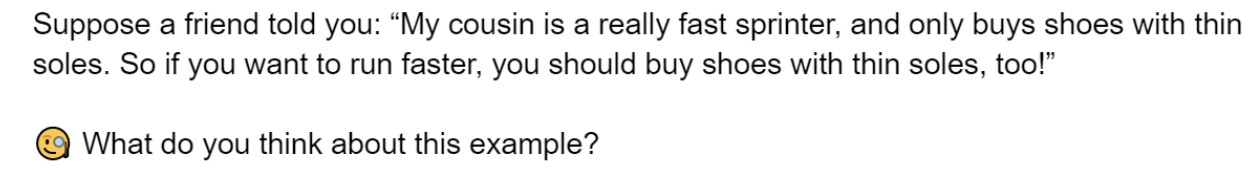
We decided to focus on only three specific manipulative tactics because we have very limited time to deliver our intervention. We picked one tactic already included in the First Draft SMS course and the pilot: misleading graphs. We chose two other tactics that are less studied in the literature and not yet in the First Draft course, because of the Partner’s above-mentioned request to investigate new tactics. 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.



In designing the intervention, we wanted to maximize participant engagement. 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). 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 expanded on the pilot and included the following additional elements in delivering our message for increased engagement (Klein, Junior & Barbosa (2015), Ziden & Rahman (2013)):

* Short and engaging, SMS style message snippets 
* Infographics (McCready, 2020).

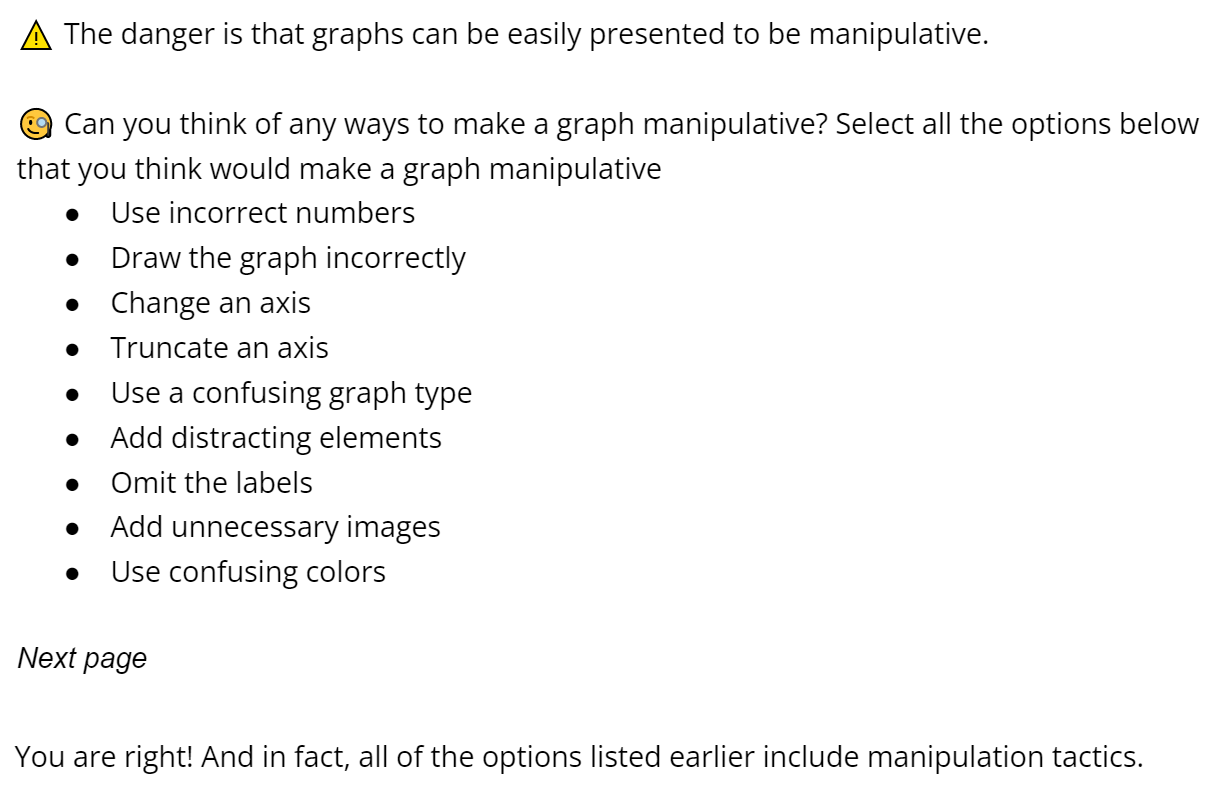


* Interactive questions to provoke thought 
* Interactive and relatable examples 

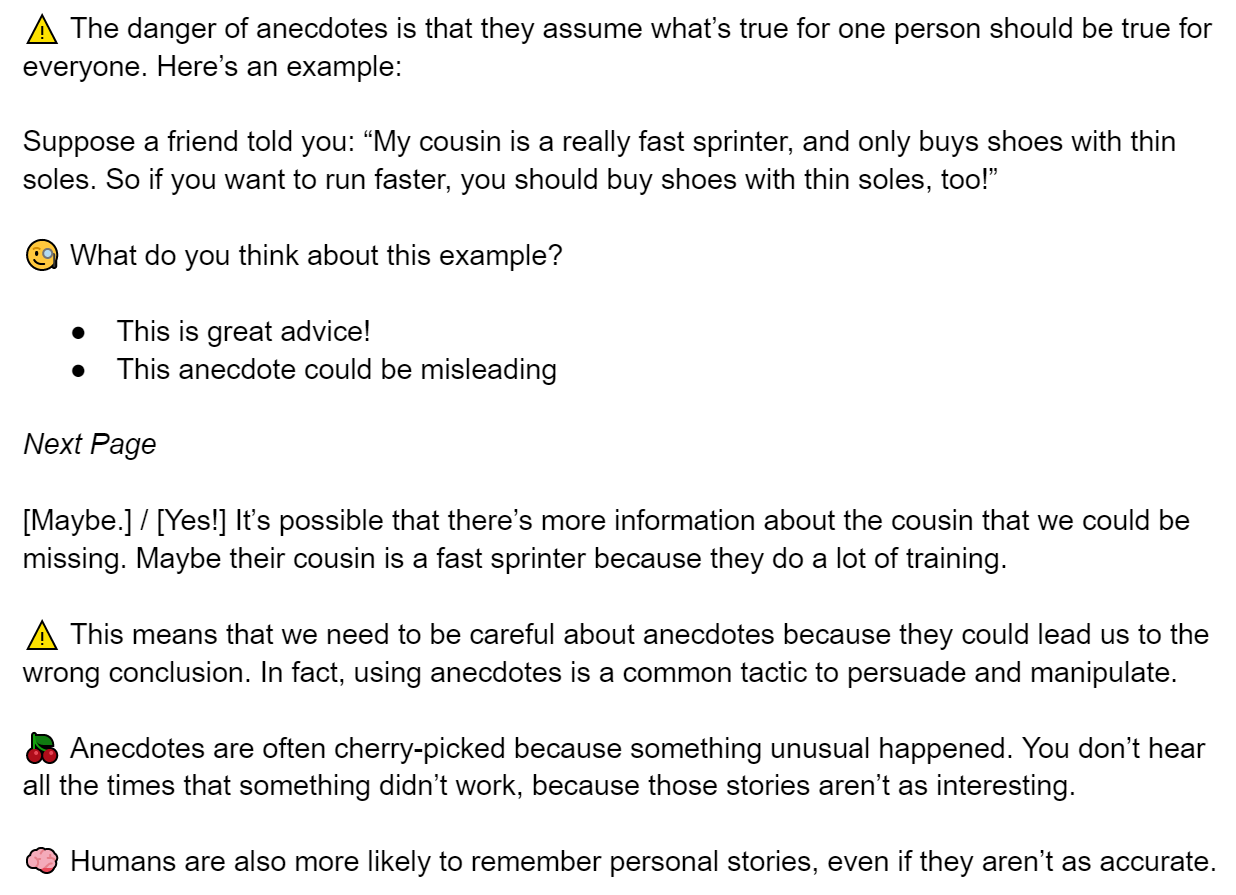
To be able to use the above interactive elements, we edited the First Draft content on manipulative graphs. Additionally, we designed novel content on false comparisons and manipulative anecdotes. We made sure that all three tactics are covered in a comparable and analogous way using the above mentioned engagement strategies.

See below some key snippets of our intervention content. The full survey script, including the full intervention, can be accessed here: <https://docs.google.com/document/d/1t88qhy2M1ysVQHT0DTT5LrQTa2_ALSI1IWRADCECPn8/edit?usp=sharing>

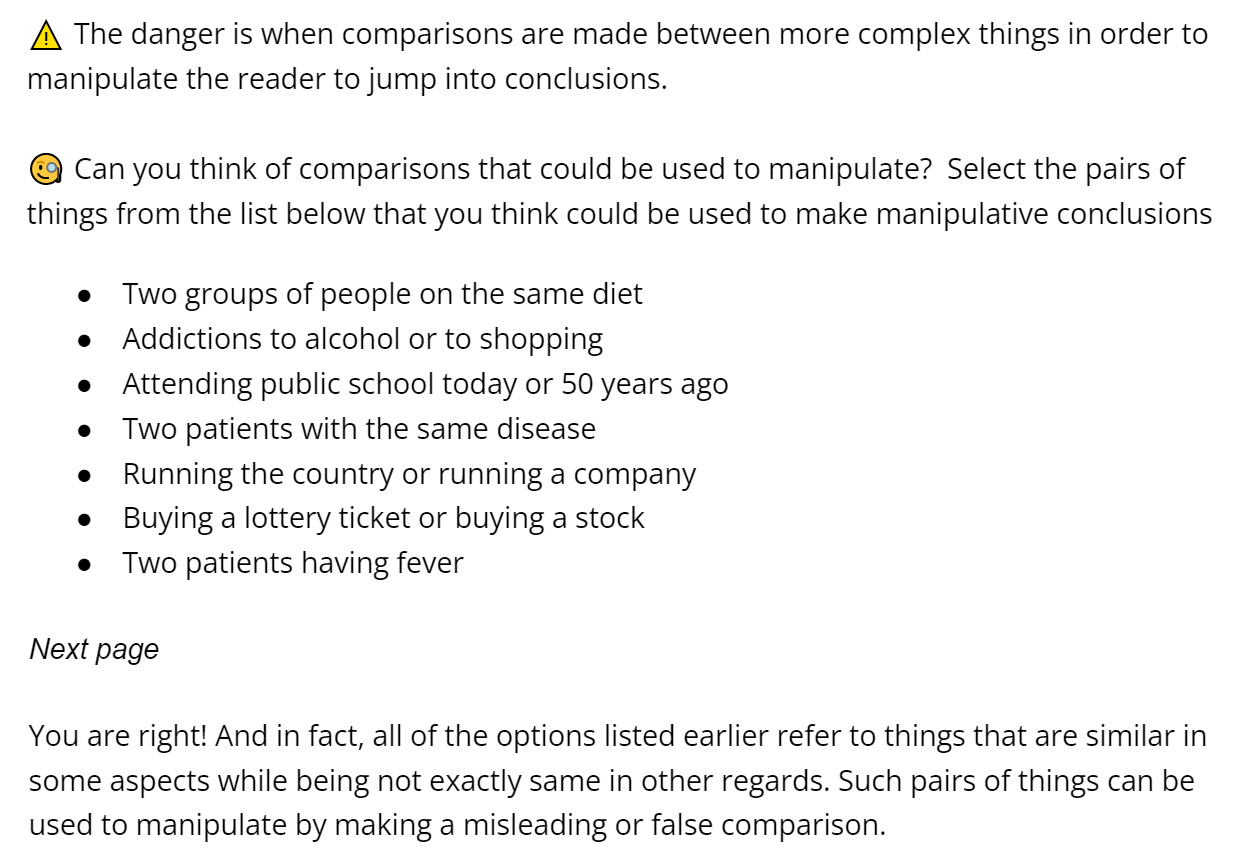
Key pieces of the misleading graphs content:



Key messages of the anecdotes content:



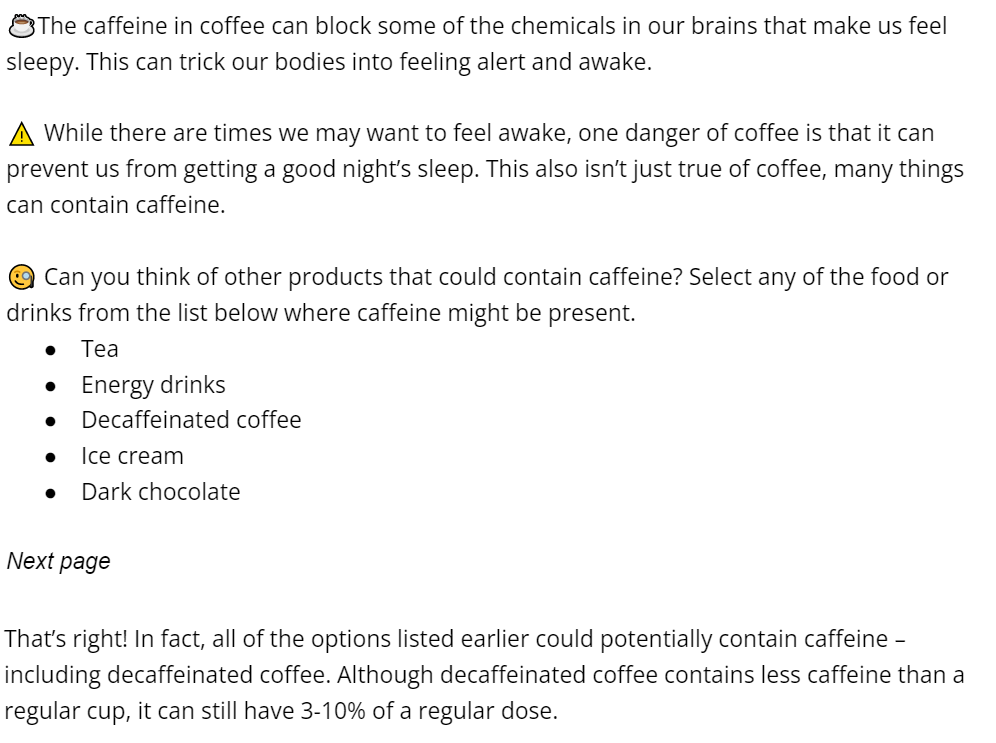
Key pieces of the false comparisons message:



## ## Control

Our final consideration in treatment design is selecting a control message to be 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. To make the control “intervention” as comparable to the treatment as possible, we focused on three key messages related to sleep: sleep quantity, caffeine, and sleep schedule. For each of the three topics, we had comparable delivery styles and lengths as for the three parts of the treatment messages.

Here is an example screenshot of the control message:



The full survey script, including the full control message, can be accessed here: <https://docs.google.com/document/d/1t88qhy2M1ysVQHT0DTT5LrQTa2_ALSI1IWRADCECPn8/edit?usp=sharing>

# # Results

## ## Summary of Insights

1. Our 6-minutes tactics focused course helped participants label manipulative posts as 12 percentage points more manipulative.
2. At the same time, we see increased skepticism for posts in general. Participants also labeled non-manipulative posts as 15.4 percentage points more manipulative. In particular, participants were more skeptical of non-manipulative graphs.
3. The rate at which participants can identify the tactics taught in our course increases (although this is only statistically significant for misleading graphs and anecdotes). However, we see drops in accuracy across all 3. This drop is due to a general increase in the frequency that participants selected tactics in the post-test. There is a (not statistically significant) increase in false positive identification, suggesting the users are not precise in their recognition of these tactics.
4. Our tactics course did not seem to change online and offline sharing behavior.
5. We found a heterogeneous treatment effect (HTE) for income only. In particular, we found that our treatment was relatively more effective for participants with lower incomes compared to participants with higher incomes. Otherwise, the treatment was consistent for participants with different ideologies and baseline misinformation susceptibility.
6. We performed additional covariate balancing and variance reduction techniques. These helped improve the significance of our estimates slightly, but not enough to change any of our initial conclusions.

## ## Recommendations for First Draft

Based on these results, we recommend that First Draft:

1. **Continue running courses to educate the public about misinformation by teaching them to recognize manipulative tactics.** The pilot study for our experiment investigated courses that focused on tactics, emotions, and a combination of the two. The pilot data suggested that the tactics course had the highest point estimate of ~3.65 amongst all the options tested (although the difference wasn’t statistically significant). The results of our experiment similarly suggest that a tactics course can help the public identify manipulative content. Relative to the control group, the treated group rated manipulative content as 12.6 percentage points more manipulative on a 6-point scale. We also found little evidence of HTE, which suggests that our course is widely effective across many different demographic groups.

There are, however, caveats to this recommendation. Future research would have to investigate the longer-term impact of treatment, and whether this helps mitigate the increased skepticism participants had for non-manipulative information. If not, the course may have to be modified. Moreover, our treatment was limited to 2 minutes per tactic because the length of our survey was constrained. First Draft may want to investigate if a longer course is more effective. These caveats are discussed in more detail in the sections on Limitations and Future Experiments below.

1. **Consider adding material about other tactics to their courses.** First Draft’s tactics course currently includes material about misleading graphs and imposter content (among other tactics). At their request, our treatment taught participants about anecdotes and false comparisons – which hadn’t previously been included in a First Draft course. Our results suggest that teaching participants about these new tactics can be just as effective as those that First Draft currently uses. In particular, relative to the control group, the treated group improved their ability to identify anecdotes and false comparisons by 17 percentage points and 12 percentage points, respectively (using the accuracy metric without a false positive penalty). This is comparable to the result for misleading graphs which increased by 19 percentage points. These tactics are all based on Cook’s (2020) FLICC taxonomy. Extrapolating from our results, other tactics in the FLICC taxonomy may be just as effective as those tested in our experiment and currently taught by First Draft. Note, however, that the same caveats discussed for the first recommendation also apply.
2. **Investigate whether our results can be replicated if the course is delivered via SMS.** First Draft’s courses are designed to be delivered via SMS in snippets over a period of 14 days. Our experiment was constrained because our course needed to be delivered to participants in one continuous sitting over the course of a ~15 minute survey (on average, our participants spent 7.02 minutes completing the treatment course). Although our course was designed to convey multiple short snippets of messages which could also be delivered via text messages, further research is needed to verify any difference in effectiveness based on different delivery formats. Using multiple text messages to spread out the information over the course of multiple days and weeks may mean that participants learn and retain the material differently.
3. **Use more interactive elements to deliver their courses.** First Draft already leverages the SMS format of their course to ask participants questions and send them automated responses. In line with findings from the literature – which suggest that these interactive elements can increase participant engagement – we tailored our course to include more thought-provoking questions and relatable examples (Roblyer & Wiencke (2003), Hake (1998)). These are discussed in more detail in the Methodology Section above. Anecdotally, the free-form comments suggest these techniques seem to have improved the overall quality of the survey. Some participants, for instance, wrote that “this survey was very different from any others I’ve taken in the past.” Others wrote that “this was the most interesting survey I’ve taken in a long time,” and that the “survey was educational and very unique.”

## ## Limitations

The conclusions we could draw from our experiment were limited by format, budget, and time constraints. These limitations are discussed in more detail below.

1. **Our results suggest participants became more skeptical of non-manipulative posts (as well as manipulative ones).** 
   * If this skepticism persists, the results could be harmful. Like misinformation, undermining the validity of legitimate information can also be damaging.
   * Some of this skepticism may be attributed to the difficulty of our test questions. For example, some of our true questions contained non-misleading anecdotes, (despite anecdotes being taught as a manipulative tactic), which may have confused participants.
   * What’s more, the long-term impacts of treatment may be different for manipulative and non-manipulative content. For instance, some of the effects we observe may be the result of priming or the experimenter demand effect. But, long-term, if participants retain information about some of the tactics but aren’t pressured by the immediacy of the course, then they may be better able to distinguish manipulative from non-manipulative content.
   * Because our experiment was limited to a single survey per participant, it was unable to measure the medium- to long-term impact of treatment. Furthermore, we didn’t measure how participants interpreted the 6-point scale they used to rate manipulativeness. This prevented us from further investigating whether they were responding to misleading and non-misleading content differently.
2. **Participants were not precise in their ability to recognize the tactics taught during treatment.**
   * As a result of treatment, participants tended to select 1.63 more tactics per question, on average. However, this increased both the number of true and false positives that they chose. This makes it difficult to measure how effective our course was at teaching participants to identify each tactic.
   * This result may suggest that treated participants became more cautious: given what they learned in treatment, they deliberately cast a wider net even if they weren’t sure if the tactics were present or not. One limitation of our treatment is that it didn’t explicitly teach participants that false positives are not desirable.
   * Even if participants did become more cautious, we can’t validate that this is what happened based on the data we collected. We didn’t ask participants about their confidence a tactic was present when they selected it.
3. **Treatment only had a minor impact on behavioral outcomes.**
   * In our design, we assumed that participants would understand that sharing manipulative information is harmful. We therefore expected that successfully teaching participants to identify manipulative content would reduce their sharing behavior. Given that our treatment effects were relatively small, it is possible that any secondary impact on sharing was even smaller and that we weren’t sufficiently powered to detect it.
   * Treatment also did not explicitly steer participants towards sharing content more or less. Moreover, there are many other factors we didn’t measure which may impact sharing behavior. For instance, the survey didn’t ask users about their intention if they said they wanted to share a post. It is therefore possible, for example, that they wanted to share a manipulative post and debunk it by explaining why it contained misinformation.
4. **Our experiment identified very few HTE.**
   * The absence of many HTEs is a strength of our treatment. It suggests that our course is effective for a wide range of demographic groups. However, the absence of HTE for ideology and misinformation susceptibility conflicts with findings in other research (e.g. Garrett & Bond (2021)).
   * For ideology, this likely reflects the fact that the posts we showed participants were deliberately politically neutral. (We also used the term manipulative information that, unlike misinformation, is not polarized.) Thus, one limitation of our study is that we don’t know the impact of treatment on politically charged content. This is significant because some disinformation actors produce content that is intentionally polarizing.

For misinformation susceptibility, this may reflect a weakness in how we measured this susceptibility. Other work has developed tailored indices to measure heterogeneity based on susceptibility (e.g. Maertens et al. (2021), Bruder et al. (2013)). We opted not to use these indices in our experiment to condense the length of the survey.

## ## Future Experiments

To address the limitations discussed above, future experiments could use several approaches:

1. **To address participants becoming more skeptical of non-manipulative posts, a future experiment could discourage false positives. Specifically, it could:**
   * Include a course module explaining why false positives can also be harmful.
   * Provide incentives (e.g. small payments), if participants can successfully identify the tactics in a post without any false positives or negatives.
   * Ask participants how confident they are that each tactic they identify in a post is actually there. This could help researchers distinguish between participants who are making emphatic mistakes, to those being cautious by casting a wider net.
   * Measure how participants are interpreting the 6-point scale used to rate manipulativeness (e.g. using an arm with a binary outcome for comparison). This could help assess if there are any non-linearities to account for.
   * Better calibrate the test questions. Some of our test questions may have unintentionally led participants astray.
2. **To decouple whether participants are learning the tactics (or if our results reflect a priming or experimenter demand effect), a future experiment could:**
   * As in point (1), provide incentives (e.g. small payments), if participants can successfully identify the tactics in a post without any false positives or negatives, and ask participants how confident they are that each tactic they identify in a post is actually there.
   * Follow up with the same participants several days or weeks after treatment to measure the long-term impact of the course.
   * Use a longer intervention and/or multiple touch points. This could involve delivering treatment via SMS over a period of 14 days, which would better approximate the true experience of a First Draft course. It could also permit the experiment to investigate if spending more than 2 minutes per tactic is more effective.
   * Provide opportunities for follow-up actions (such as allowing participants to fact check specific posts). This would also better reflect the resources participants would have access to in the real world.
   * Observe real-world behavior after the experiment, for example by installing an app on participants’ phones (with their consent). This may also be beneficial for First Draft given that influencing behavior on social media is ultimately their outcome of concern.
3. **To more actively shape behavioral outcomes, future experiments could:** 
   * Contain a module explaining how participants should address manipulative posts they encounter.
   * Ask participants about their motivation for taking actions.
   * Match topics in the posts used for the pre- and post-tests to account for participants’ interests.
   * Explain the potential negative consequences of sharing misinformation.
4. **To compare the HTE for our course to those observed in other literature, future experiments could:**
   * Include some politically-polarizing posts.
   * Measure misinformation susceptibility using several indices developed and validated elsewhere (e.g. in Maertens et al. (2021) and Bruder et al. (2013)).
   * Increase sample size or the length of the intervention to have more power for identifying potentially small HTE.

# # References

* Bjälkebring, P., & Peters, E. (2021). Money matters (especially if you are good at math): Numeracy, verbal intelligence, education, and income in satisfaction judgments. PloS one, 16(11), e0259331.
* Bruder, M., Haffke, P., Neave, N., Nouripanah, N., & Imhoff, R. (2013). Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy Mentality Questionnaire. Frontiers in psychology, 4, 225.
* Kricorian, K., Civen, R., & Equils, O. (2022). COVID-19 vaccine hesitancy: Misinformation and perceptions of vaccine safety. Human Vaccines & Immunotherapeutics, 18(1), 1950504.
* Cook, J. (2020). Deconstructing climate science denial. In Research Handbook on Communicating Climate Change. Edward Elgar Publishing.
* Pan, W., Liu, D., & Fang, J. (2021). An examination of factors contributing to the acceptance of online health misinformation. Frontiers in psychology, 12, 524.
* Garrett, R. K., & Bond, R. M. (2021). Conservatives’ susceptibility to political misperceptions. Science Advances, 7(23), eabf1234.
* 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.
* Maertens, R., Götz, F., Schneider, C. R., Roozenbeek, J., Kerr, J. R., Stieger, S., ... & van der Linden, S. (2021). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment.
* Rapp, D. N., & Salovich, N. A. (2018). Can’t we just disregard fake news? The consequences of exposure to inaccurate information. Policy Insights from the Behavioral and Brain Sciences, 5(2), 232-239.
* 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.