Twitter Accuracy Nudge Revision (#14535)

Created: 10/01/2018 07:52 PM (PT)
Shared: 11/08/2019 10:11 AM (PT)

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?
Here we test the hypothesis that sending Twitter users a direct message (DM) asking them to rate the accuracy of a single non-partisan headline will improve the quality of the news content they subsequent share on Twitter.

3) Describe the key dependent variable(s) specifying how they will be measured.
We will define the quality of a user's tweets using a predefined list of 60 domains which have each been given a quality rating (between 0 and 1) by professional fact-checkers (from Pennycook & Rand 2018). For each user, we will extract all statuses (retweets and tweets) from the user's twitter account over the relevant timeframe that contain links to one of the 60 domains in the list. Our analysis will be conducted at the level of the tweet, and the fact-checker quality rating of the domain linked to in the tweet will be our dependent variable. Our primary analysis will examine the 24 hours after we send the user a DM; secondary analyses will examine subsequent days to assess persistence.

4) How many and which conditions will participants be assigned to?
In order to send a Twitter user a DM, the user must follow your account. Thus, a first stage of the experiment is to create the "subject pool" of message-able users. To do so, we created a set of seven accounts all named CookingBot (we created multiple accounts to address Twitter rate-limits on # of follows and # of DMs per day), and had each account follow a large number of twitter users who have retweeted at least one link to either Breitbart or Infowars (two well-known sites that post a large amount of misinformation). Those users who followed our account back form our subject pool.

Within this subject pool, each subject will be assigned to one of two conditions, control or treatment. Subjects in the treatment will receive a DM from one of our accounts which thanks the user for following, shows a true (but ambiguous) non-political article, and asks the user to rate its accuracy on a 4-point scale. Subjects in the control receive no message.
Subjects will be randomly assigned to condition, stratified on count of links to one of the 60 sites in our list in the 14 days before the experiment (median split); average quality of links (re)tweeted in the 14 days prior to the beginning of the experiment (using median split); and ideology (using median split, and estimating ideology by the method of Barbera et al 2015). DMs will be sent on one of 3 sequential days (to avoid Twitter DM rate limits); thus subjects are also be randomly assigned to a treatment date.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Our primary analyses will be as follows. For each user in the treatment condition, we will examine all of their statuses containing links to one of the 60 websites in our list in the 24 hrs after receiving the DM; and compare this to all of the statues in that same 24 hr time window of all users who had not received a DM (all control subjects, and all treatment subjects whose DM time was after that 24 hr window). We will do so using a linear regression with robust standard errors clustered on user, predicting link quality, with a treatment dummy (0=not received DM, 1=received DM) as the independent variable. To calculate an overall DM effect size, we will construct a weighted average of the individual user DM effect sizes. We will do this in two ways: by weighting each treatment user equally, and by weighting each treatment user by their number of tweets in the relevant 24 hr period (ie by the amount of data they provide). In both cases, we will then calculate 95% confidence intervals, as well as a p-value for a positive effect, using bootstrapping (randomly select user status histories – ie clustering at the level of the user).

Secondarily, we will run the same analyses using data from subsequent days to examine how the treatment effect changes over time.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We will exclude participants who did not tweet links to any of the 60 sites in our list in the two weeks prior to the experiment; who could not be given an ideology score; who could not be given a botornotscore; or who had a botornotscore above 0.5.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
Our sample will consistent of 2211 users (those remaining after exclusions, and 2 users who were accidentally DMed twice).

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will replicate our main analysis restricting just to statuses that were retweets with no accompanying text.

We will use machine learning approaches to explore potential moderators of the treatment effect.

The data for this experiment have already been collected, but have not yet been examined/analyzed at all. Thus this pre-registration is still valid. This situation arose because Twitter disabled its DM API during the 2nd day of our data collection. As a result, we had to DM users manually, and thus DM times were spread over a long duration. This made our original preregistered analysis plan invalid, and led us to the above plan. We also note that our original randomization scheme placed 1107 users in the control and 1104 in the treatment. However, due to the API issues, DMs were definitely not sent to 186 users assigned to treatment (there was no DM timestamp recorded for these users), and may not have been sent to another 91 users assigned to treatment (these users have a DM timestamp, but no DM appears in our account's DM log). Our main analysis will not include treatment effects for the 186 users with no timestamps (without a timestamp it is impossible to calculate an effect) but will include the 91 users with timestamps. Secondarily, we will see how the effect estimates change if these 91 users are classified as untreated instead of treated (ie it is assumed no DM was ever sent to them).