PREREGISTRATION

Background and explanation of rationale

How can we most effectively counter false information in politics, especially online? Social science still has few answers to these questions (Flynn, Nyhan, and Reifler 2017). However, the spread of "fake news" online during the 2016 election highlighted the necessity of developing more effective responses to misinformation. Facebook has formed a new partnership with fact-checkers to counter the spread of false information on its platform, but little is known about the effectiveness of the approach the company has undertaken or what might be more effective. We will therefore conduct an experiment testing two approaches to countering belief in misinformation on social media – an advance warning, which Bolsen and Druckman (2015) find is most effective in scientific controversies, and a Facebook-style banner identifying claims that fact-checkers rate as inaccurate, which Pennycook et al. find is effective (2017).

Our study makes several contributions. First, we extend the advance warning approach to the context of false headlines online. Second, Pennycook et al. did not explore whether Facebook's "disputed" banners were as effective at dissuading belief in false headlines as a direct refutation identifying a false article as such. This hypothesis conforms to our theoretical understanding of fact-checks. For instance, Fridkin et al. (2015) finds that fact-checks which directly and negatively question the truth of a negative political ad "influence people's assessments of the accuracy, usefulness, and tone of negative political ads" (145). Third, we consider whether the presence of an advance warning strengthens the effect of "disputed" or "false" flags. Finally, we consider the effects of these treatments on intended liking and sharing behavior on Facebook as well as possible spillovers to the perceived accuracy of both true articles and unflagged false articles.

What are the hypotheses to be tested/quantities of interest to be estimated?

We plan to test the following four experimental hypotheses.

Hypotheses:

H1: Exposure to a general warning about misleading articles will reduce the perceived accuracy of false headlines (relative to a no-warning condition).

H2a: The presence of a "disputed" flag under false headlines will reduce their perceived accuracy (relative to a no-flag condition).
H2b: The presence of a "false" flag under false headlines will reduce their perceived accuracy (relative to a no-flag condition).
H2c: The presence of a "false" flag under false headlines will reduce their perceived accuracy relative to when a "disputed" flag appears under them.

H3: Exposure to a general warning about misleading articles will increase the negative effects of "disputed" or "false" flags on the perceived accuracy of false news headlines.

H4a: The effect of a "disputed" flag on perceived accuracy (versus a headline with no flag) will be reduced for politically congenial information versus uncongenial information.
H4b: The effect of a "false" flag on perceived accuracy (versus a headline with no flag) will be reduced for politically congenial information versus uncongenial information.

We will plan to explore the following research questions, which raise important issues for which we do not have strong theoretical expectations:

RQ1: Will a warning about misleading articles or the presence of a "disputed" or "false" flag on a headline affect the likelihood that people will say they would like or share it?

RQ2: Will the effect of a warning about misleading articles on the perceived accuracy of subsequent headlines (versus no warning) vary between congenial information and uncongenial information?

RQ3: Will the presence of "disputed" or "false" flags affect the perceived accuracy of unflagged false (RQ3a) or true (RQ3b) news (relative to a no-flag condition)?

RQ4: Will a warning about misleading articles reduce the perceived accuracy of true information (relative to a no-warning condition)?

[Note: H1-H4 and RQ1-RQ3a concern the effects of the experimental manipulations on the perceived accuracy and likelihood of "liking" or sharing false headlines/claims only. RQ3b-RQ4 consider the effects of the experimental manipulations on the perceived accuracy of true headlines/claims.]

How will these hypotheses be tested?

[All of the survey items and the experimental protocol are attached below.]

Eligibility and exclusion criteria for participants

Participants will be United States residents age 18 and over recruited on the Amazon Mechanical Turk online marketplace. All Turkers are eligible to participate in this study who meet the specified qualifications and did not take part in an earlier pilot study. The sample size will be approximately 2500-3500 – data collection will continue until all funds allocated to the project are exhausted. Researchers have no role in selecting the participants after listing the project on Mechanical Turk.

Randomization approach

We will use a between-subjects design in which respondents are randomly assigned to one of seven conditions by the Qualtrics online survey platform (p=1/7 for each):

-Pure control (no warnings or headlines shown)
-Warning about false and misleading stories, no flags on false headlines
-Warning about false and misleading stories, disputed flags on false headlines
-Warning about false and misleading stories, false flags on false headlines
-No warning about false and misleading stories, no flags on false headlines
-No warning about false and misleading stories, disputed flags on false headlines
-No warning about false and misleading stories, false flags on false headlines

Data collection and blinding

Data will be collected on the Qualtrics online survey platform. There may be some online discussion among Mechanical Turk workers about the details of our survey. This cannot be prevented and we hope that these participants preserve the integrity of the research.

Primary and secondary outcome measures

Our principal outcome measures are the perceived accuracy of the relevant set of false claims that were shown in headlines to respondents not assigned to the pure control condition (the set that is relevant varies by hypothesis; see below). The wording of these questions appears below. (Note: We will also conduct exploratory analysis of individual outcome questions.)

To the best of your knowledge, how accurate is the claim that Trump is bringing back the draft?
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

To the best of your knowledge, how accurate is the claim that Trump plagiarized the Bee Movie for his inaugural speech?
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

To the best of your knowledge, how accurate is the claim that Republican Congressman Jason Chaffetz was blackmailed by the Kremlin?

-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

To the best of your knowledge, how accurate is the claim that a Donald Trump protester was paid $3,500 to protest Trump's rally?
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

To the best of your knowledge, how accurate is the claim that Donald Trump sent his own plane to transport 200 stranded Marines?
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

To the best of your knowledge, how accurate is the claim that an FBI agent suspected in Hillary Clinton's email leaks was found dead in an apparent murder-suicide?
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

Secondary outcome measures:

- The average likelihood of “liking” or sharing a false article on Facebook (among respondents who do not answer “never” when asked how often they use Facebook)
- The average perceived accuracy of true articles

The wording of these questions appears below.

Liking/sharing questions:

After the relevant perceived accuracy question, the following questions appear below each headline for respondents not assigned to the pure control condition who did not answer “never” when asked how often they use Facebook:

How likely would you be to "like" this story on Facebook?
-Very likely (4)
-Somewhat likely (3)
-Not very likely (2)

-Not at all likely (1)

How likely would you be to share this story on Facebook?
-Very likely (4)
-Somewhat likely (3)
-Not very likely (2)
-Not at all likely (1)

True article outcome measures:

To the best of your knowledge, how accurate is the claim that Trump questioned why the U.S. Civil War had to happen?
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

To the best of your knowledge, how accurate is the claim that Trump ordered airstrikes in Syria after a chemical attack?
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

To the best of your knowledge, how accurate is the claim that Neil Gorsuch was confirmed to the Supreme Court?
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

Statistical analyses

All results will be estimated as stacked question-level data using OLS with robust standard errors clustered by respondent and respondent and question fixed effects. These results will be verified for robustness using appropriate GLM estimators (see below). We exclude control variables in our experimental analyses (see below). Unless otherwise noted, all experimental treatment effects will be estimated as intent to treat effects.

Main effects (including interactions between experimental conditions)

$Y = b0 + b1*nocorrection + b2*disputed + b3*false + b4*warning + b5*disputedXwarning + b6*falseXwarning$

where nocorrection=1 if respondents were assigned to the false headlines with no flags condition and 0 otherwise; disputed=1 if respondents were assigned to the false headlines with "disputed" flags condition and 0 otherwise; false=1 if respondents were assigned to the false headlines with "false" flags condition and 0 otherwise; and warning=1 if respondents were assigned to the condition with a warning about misleading articles before viewing headlines and 0 otherwise. The excluded category are pure controls. (For expositional clarity, we may estimate an equivalent model with pure controls excluded and the no correction / no warning condition as the baseline category and report this model in an appendix.)

For H1-H3, the outcome measure is the perceived accuracy of false headlines. Responses from respondents in the disputed and false conditions for false headlines that were randomly assigned not to receive a flag are excluded (see RQ3a below).

We will test H1 by computing b4-b1 (the effect of warning exposure versus the headlines/no warnings condition).
We will test H2a by computing b2-b1 (the effect of "disputed" flag exposure versus the headlines/no warnings condition).
We will test H2b by computing b3-b1 (the effect of "false" flag exposure versus the headlines/no warnings condition).
We will test H2c by computing b3-b2 (the effect of "false" flag exposure versus "disputed" flag exposure).
We will test H3 by estimating b5 (how effect of disputed flag *changes* given warning) and b6 (how effect of false flag *changes* given warning). We will also compute the relevant marginal effects using b2-b1+b5 (effect of disputed given warning) and b3-b1+b6 (effect of false given warning) and contrast these with the marginal effects absent warning (b2-b1 and b3-b1, respectively).

For RQ1, the outcome measure will instead be respondents' reported average likelihood of "liking" an article or sharing it on Facebook among those who do not respond "never" when asked how often they use Facebook. Responses from respondents in the disputed and false conditions for false headlines that were randomly assigned not to receive a flag are again excluded. To explore RQ1, we compute b4-b1 (the effect of warning exposure versus the headlines/no warnings condition), b2-b1 (the effect of "disputed" flag exposure versus the headlines/no warnings condition), and b3-b1 (the effect of "false" flag exposure versus the headlines/no warnings condition).

For RQ3a, the outcome measure will be the perceived accuracy of false articles that were not flagged as such. In this case, we will exclude flagged false articles for respondents in the disputed and false conditions as well as responses in which respondents who were randomly assigned to those conditions had not previously seen a disputed or false flag on a headline. We will then estimate the model above and compute b2-b1 (the effect of previous "disputed" flag exposure versus the headlines/no

warnings condition), and b3-b1 (the effect of previous “false” flag exposure versus the headlines/no warnings condition).

For RQ3b and R4, the outcome measure will instead be the perceived accuracy of true articles. To test for spillover effects, we will only include responses from the disputed and false conditions in which respondents assigned to those conditions had previously seen a disputed or false flag on a headline. We will then estimate the model above and compute b2-b1 (RQ3b: the effect of “disputed” flag exposure versus the headlines/no warnings condition), b3-b1 (RQ3b: the effect of “false” flag exposure versus the headlines/no warnings condition), and b4-b1 (RQ4: the effect of warning exposure versus the headlines/no warnings condition).

Interactions with directional preference measure(s)

Y = b0 + b1*disputed + b2*false + b3*warning + b4*warningXdisputed + b5*warningXfalse + b6*trump_congenial + b7*disputedXtrump_congenial + b8*falseXtrump_congenial + b9*warningXtrump_congenial + b10*warningXdisputedXtrump_congenial + b11*warningXfalseXtrump_congenial

where trump_congenial = 1 if headline is anti-Trump and respondent disapproves of Trump or headline is pro-Trump and respondent approves of Trump and trump_congenial = 0 if headline is anti-Trump and respondent approves of Trump or headline is pro-Trump and respondent disapproves of Trump. (For expositional clarity, these models will exclude pure independents and respondents assigned to the pure control condition – models including all respondents that include interactions between trump_congenial and trump_uncongenial and each experimental condition will be reported in an appendix.)

We will test H4a by estimating b7 and H4b by estimating b8. We will also compute the relevant marginal effects of disputed (b1+b7) and false (b2+b8) and contrasting them with the relevant marginal effects for uncongenial headlines (b1 and b2, respectively). Similarly, we will test RQ2 by estimating b9 and contrasting the relevant marginal effect (b3+b9) with the relevant marginal effect for an uncongenial headline (b3).

Notes:
-We will compute and report appropriate auxiliary quantities from our models to test the hypotheses of interest, including marginal effects appropriate to test the hypotheses of interest from the models including interaction terms, treatment effects by subgroup, and differences in marginal effects between subgroups.
-In some cases, we may present treatment effects estimated on different subsets of the data for expositional clarity. If so, we will verify that we can reject the null of no difference in treatment effects in a more complex interactive model reported in an appendix when possible.

-For interaction terms, scales, and moderators, if results are consistent using a median/tercile split or indicators rather than a continuous scale, we may present the latter in the main text for ease of exposition and include the continuous scale results in an appendix. We will also use tercile indicators to test whether a linearity assumption holds for any interactions with continuous moderators per Hainmueller et al (N.d.) and replace any continuous interactions in our models with them if it does not.
-Don't know responses will be considered missing data for the factual belief outcome measures.
-We will compute and report summary statistics for our samples. We will also collect and may report response timing data as a proxy for respondent attention.
-The order of hypotheses and analyses in the final manuscript may be altered for expositional clarity.
-Where applicable, regression results for binary dependent variables will be verified for robustness using probit. Regression results for individual ordered dependent variables will be verified for robustness using ordered probit.
-We may estimate the experimental models described above with a standard set of covariates (indicators for gender, age groups, non-white respondents, respondents with a four-year college degree, and scores on a standard political knowledge scale) if including those has a substantively important effect on the precision of our treatment effect estimates. In that case, however, both models will be reported.