EGAP PREREGISTRATION

Title of Study

Effects of fake news tips and fake news exposure (fall 2018 replication and extension)

Authors

Andrew Guess, Princeton University (aguess@princeton.edu)

Benjamin Lyons, University of Exeter (B.Lyons@exeter.ac.uk)

Jacob Montgomery, Washington University in St. Louis (jacob.montgomery@wustl.edu)

Brendan Nyhan, University of Michigan (bnyhan@umich.edu)

Jason Reifler, University of Exeter (J.Reifler@exeter.ac.uk)

Acknowledgements

We gratefully acknowledge support from Democracy Fund and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 682758).

Is one of the study authors a university faculty member?

Yes

Is this Registration Prospective or Retrospective?

Registration prior to researcher access to outcome data

Is this an experimental study?

Yes

Date of start of study:

10/19/2018

Gate date:

May 3, 2020

Was this design presented at an EGAP meeting?

No

Is there a pre-analysis plan associated with this registration?

Yes

Background and explanation of rationale

Despite widespread hype about the role of “fake news” in the 2016 election (e.g., Albright, 2017), actual consumption of fake news was limited and concentrated among a small set of people with strong tendencies toward selective exposure (Guess et al., 2018). Much remains to be learned about fake news, however. In particular, little is known about what effects it has on people’s beliefs and attitudes or how we might best counteract it.

One approach to countering fake news is to increase media literacy by training people to recognize fake news online (Kiely and Robertson, 2016; Caulfield, 2017; First Draft, 2017; Jang & Kim, 2018; Tandoc et al., 2018). In the first wave of this study, we thus examine whether a news literacy intervention - showing people tips to help them spot fake news - can improve respondents’ ability to distinguish between real and fake news (see also Blair et al. N.d., who find that fact-check banners help reduce the perceived accuracy of fake news but that general warnings about fake news decrease the perceived accuracy of both fake and real news). We also examine factors that may condition the effects of the intervention.

One apparent danger of fake news is that mere exposure may increase its subsequent believability (Pennycook et al., 2018). We test this by manipulating prior exposure to fake news and evaluating whether articles that have been seen before are seen as more accurate. We also examine whether exposure to a news literacy intervention can limit this effect.

Research on partisan media effects suggests that fake news exposure may also increase affective polarization and willingness to take political action (Garrett et al, 2014; Lau et al., 2017; Suhay et al., 2017; Tsfati & Nir, 2017). An experiment embedded in the second wave of this study therefore examines whether exposure to pro- or counter-attitudinal fake news changes levels of affective polarization, affect toward the media, and intent to engage in political action. Finally, we explore the potential for individual differences in fake news effects.

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

Note: We have amended our pre-registration to reflect a programming error in the survey experiment conducted in this study (we discovered it after examining the data for 20181106AE, which was programmed nearly identically and contained the same error). Our news tips intervention was programmed such that all respondents were exposed. Accordingly, we have removed any hypotheses and models examining the effects of news tips from the set of pre-registered analyses below. For full transparency, we have used strikethrough formatting to show which hypotheses and analyses we have removed.

Wave 1

We plan to test the following hypotheses using survey data from Wave 1 and behavioral data on the web traffic patterns of respondents:

A. Observational hypotheses

H-A1) People with the strongest overall tendencies toward selective exposure will be the most likely to consume fake news and consume the most on average. (This hypothesis tests if the Guess et al. results replicate in this sample.)

H-A2) People who consume fake news will be more likely to believe it is accurate than those who do not consume fake news (H-A2a). This relationship will be stronger for pro-attitudinal fake news belief than for counter-attitudinal fake news belief (H-A2b) and for people who are relatively less skilled at analytical reasoning (H-A2c). We also expect that people who consume fake news will be less likely to successfully distinguish between true and false headlines (H-A2d).

H-A3) People who consume fake news will be more likely to hold topical misperceptions than those who do not consume fake news (H-A3a). This relationship will be stronger for pro-attitudinal misperceptions than for counter-attitudinal misperceptions (H-A3b) and for people who are relatively less skilled at analytical reasoning (H-A3c). People who consume fake news will be less likely to successfully distinguish between true and false topical statements (H-A3d).

We will provide the first test of the general association between exposure to fake news sites and belief in both fake news headlines and topical misperceptions. This relationship is likely to vary by whether the content of the fake news is consistent with respondents' political viewpoint. We will examine fake news exposure's relationship with perceived accuracy of related, true, topical statements as a baseline. (Note: We cannot prove that any such relationship is causal but can evaluate whether respondents who view fake news indicate that they believe in the fake news headlines or topical misperceptions.)

Additionally, some research is suggestive that older respondents consume more fake news (Guess et al n.d.) and are less able to separate factual statements from opinion statements (Gottfried and Greico 2018). To more directly assess whether age plays a role in how people evaluate news, we propose the following RQ-A1: Is there a relationship between respondent age and perceived accuracy of fake news?

B. Main effects of news tips intervention

Exposure to tips for spotting fake news should do the following:

H-B1a) The news tips intervention will decrease the perceived accuracy of fake news. (It should help people identify clues that the article is dubious.)

H-B1b) The news tips intervention will increase the perceived accuracy of real news. (It should help people identify clues of credibility and/or generate a contrast effect with fake news articles.)

H-B1c) The news tips intervention will increase people's ability to successfully distinguish between real and fake news. (This implication follows from H-B1a and H-B1b.)

H-B1d) The news tips intervention will decrease the perceived accuracy of hyper-partisan news.

RQ-B1) Does exposure to the news tips intervention affect subsequent information consumption?

We will test whether respondents assigned to the news tips condition are less likely to visit fake news websites than those assigned to control (RQ-B1a). We will also test whether respondents assigned to the news tips condition are more likely to visit fact-checking websites (RQ-B1b), and mainstream news websites (RQ-B1c) those assigned to control.

RQ-B2) What effects will the news tips intervention have on intention to share fake, real, or hyper-partisan news?

Respondents should presumably be less likely to share fake and hyper-partisan news and more likely to share real news, but Pennycook and Rand find that prior exposure to an accuracy question - the same measurement approach used in our design - suppresses self-reported sharing intentions. The effects of the intervention in this context are therefore unclear.

C. Heterogeneous treatment effects

H-C1) The effect of the news tips intervention will be reduced for pro-attitudinal content compared to counter-attitudinal content. (Based on prior findings showing greater vulnerability to pro-attitudinal misinformation (e.g., Nyhan and Reifler 2010), we expect that respondents will improve less at distinguishing false from real news in response to the intervention when the content in question is consistent with their predispositions; see also, e.g., Metzger, Hartsell, and Flanagin 2015.)

In addition, we will also conduct exploratory analysis of other potential moderators of the effect of news tips on the perceived accuracy of real and fake news: Cognitive Reflection Test scores (per Pennycook and Rand 2018, but see Kahan 2018), trust in and feelings toward the media, feelings toward Trump, conspiracy predispositions, political interest and knowledge, and pre-treatment visits to fake news sites and fact-checking sites. Given the risk of false positives, we will control the false discovery rate with the Benjamini-Hochberg procedure. These analyses will be limited to the appendix or supplementary materials, but if any positive findings replicate in future studies, we may then use these data and analyses in the main text of a paper.

Waves 1-2

D) Effects of prior exposure to fake news and news tips intervention

We also examine whether exposure to either fake news or the news tips intervention have lingering effects over time. We replicate Pennycook et al. (2018), who show a single exposure to fake news increases subsequent perceptions of accuracy.

H-D1. Randomized exposure to a fake news story in Wave 1 increases the perceived accuracy of those stories in Wave 2.

H-D2. Randomized exposure to a real news story in Wave 1 increases the perceived accuracy of those stories in Wave 2.

H-D3. Randomized exposure to a hyper-partisan news story in Wave 1 increases the perceived accuracy of those stories in Wave 2.

~~We also test whether exposure to the news tips intervention changes this effect.~~

~~RQ-D1. Does exposure to the news tips intervention affect perceived accuracy of fake news story repeated in Wave 2?~~

~~RQ-D2-D5. Are any main effects of the news tips intervention on belief in real news (RQ-D2), fake news (RQ-D3), the ability of people to distinguish between real and fake news (RQ-D4), and on belief in hyperpartisan news (RQ-D5) measurable in wave 2?~~

Wave 2

E) Main effects of fake news exposure

H-E1a. Exposure to pro-attitudinal fake news will increase affective polarization. (Exposure to online criticism of the other party has been found to increase affective polarization (Suhay et al. 2017). Related evidence finds that partisan media can potentially also increase ideological polarization, although the extent of these effects and the groups that are affected remains an ongoing area of research (e.g., Arceneaux and Johnson 2013; Levendusky 2013; see Iyengar et al. N.d. and Prior 2013 for reviews).)

H-E1b. Exposure to pro-attitudinal fake news will increase negative feelings toward the media. (Ladd 2011 finds that exposure to elite media criticism and tabloid-style news increase media distrust, as does talk radio exposure among conservatives. During the 2016 election, fake news often amplified attacks on the mainstream media and used tabloid- and talk radio-style approaches.)

H-E2/E3. Exposure to pro-attitudinal fake news will increase intent to vote (H-E2) and intent to take political action (H-E3). Negative affect toward the other party, which much fake news promotes, is increasingly associated with political participation (Iyengar and Krupenkin 2018). Similarly, though advertising does not seem to affect net turnout levels, changes in the partisan balance of advertising affect partisan vote shares (Spenkuch and Toniatti, 2018). Finally,

exposure to Fox News was found to particularly increase vote intention among Republicans and independents (Hopkins and Ladd 2014).

H-E4/E5. Exposure to fake news will increase belief in the claims in the articles in question (H-E4), especially among respondents for whom those articles are pro-attitudinal (H-E5).

RQ-E1. Commentators have suggested that attacks on George Soros appeal to and promote anti-Semitic views (e.g., Ackerman 2018). We therefore examine whether exposure to a fake news article attributing the caravan to him promotes more negative sentiment toward Jews overall (RQ-E1a) or among Republicans for whom the article is pro-attitudinal (RQ-E1b).

F) Heterogeneous treatment effects

RQ-F1-4. What effect does counter-attitudinal fake news exposure have on affective polarization (RQ-F1), affect toward the media (RQ-F2), intent to vote (RQ-F3), or intent to take political action (RQ-F4)? (Research and commentary has largely focused on the effect of pro-attitudinal fake news exposure. Theoretical expectations about the effects of counter-attitudinal fake news exposure are less clear. It could have a persuasive effect as with Fox, which appears to have increased Republican vote share in part through persuasion (though conversion rates were low; see Hopkins and Ladd 2014 and Martin and Yurukoglu 2017). Alternatively, it could generate backlash, as persuasion efforts frequently fail (e.g., Bailey, Hopkins, and Rogers 2016). These effects could apply to both attitudes (affective polarization) and behavioral intentions (intent to vote or take political action).

We will also conduct exploratory analyses of potential moderators of the effect of pro-attitudinal fake news on affective polarization, intent to vote, or intent to take political action: trust in and feelings toward the media, feelings toward Trump (entered as a linear term and with indicators for terciles or quartiles), conspiracy predispositions, political interest and knowledge, and pre-treatment visits to fake news sites and fact-checking sites. In addition, we will conduct exploratory analyses of Hispanic/Latino-white differences in feeling thermometer scores as a moderator of fake news exposure effects for the immigrant caravan stimulus and Christian-Muslim differences in feeling thermometer scores as a moderator of fake news exposure effects for the Khashoggi stimulus in wave 2. As we describe above for wave 1, we will control the false discovery rate with the Benjamini-Hochberg procedure given the risk of false positives. These analyses will be limited to the appendix or supplementary materials, but if any positive findings replicate in future studies, we may then use these data and analyses in the main text of a paper.

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 are YouGov panel members in the U.S. who consented to participate in an online study (YouGov determines the specific eligibility and exclusion criteria for their panel). Researchers have no role in selecting the participants.

Randomization approach (wave 1)

~~We will use a between-subjects design in which respondents are randomly assigned to exposure to fake news tips from Facebook (randomized with p=.5 via the YouGov platform).~~

8 news articles for evaluation will be displayed according to the following logic:

[16 possible articles: 4 pro-D real news articles (2 from low-prominence sources and 2 from high-prominence sources), 4 pro-R real news articles (2 from low-prominence sources and 2 from high-prominence sources), 2 pro-D fake news articles, 2 pro-R fake news articles, 2 pro-D hyperpartisan sources, 2 pro-R hyperpartisan sources]

[show 8 articles from these sets in random order per algorithm below]

Show 4 of 8 fake or hyperpartisan article previews: 1 pro-D hyper {random from 2}, 1 pro-D fake {random from 2}, 1 pro-R hyper {random from 2}, 1 pro-R fake {random from 2}.

Each article's slant is listed in the survey document below.

In wave 2, all participants will subsequently be exposed to all 16 possible articles from wave 1, thus providing perceptions of accuracy for repeated and non-repeated articles.

Randomization approach (wave 2)

We will use a between-subjects design in which respondents are randomly assigned to exposure to fake news/hyper-partisan article via the YouGov platform (p=1/3 pro-D fake news article, 1/3 pro-R fake news article, 1/3 placebo).

Data collection and blinding

Data will be collected by YouGov via the survey waves described above and anonymized web traffic data from respondents (see Guess et al. for details on the YouGov Pulse panel, though the tracking technology and panel have changed since 2016).

Primary and secondary outcome measures

Waves 1 and 2 (survey)

Our primary outcome measures in wave 1 are the perceived accuracy of the claims that were shown in headlines to respondents (from either real, fake, or hyperpartisan news sites). These are also among the primary outcome measures in wave 2. The wording of these questions appears below.

To the best of your knowledge, how accurate is the claim in the above headline?

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

Respondents' ability to distinguish between real and fake will be measured as the mean perceived accuracy of real headlines minus the mean perceived accuracy of fake headlines.

(See the attached instrument, which indicates which articles are real, fake, or hyperpartisan, and which are coded as pro-Republican versus pro-Democrat. We will assess the robustness of the mean(real)-mean(fake) measure to violations of a unidimensionality assumption and/or differential item difficulty and substitute a more complex measure using question fixed effects, IRT, etc. if appropriate.)

A secondary outcome measure is self-reported intention to "like" or "share" an article on Facebook:

[If How frequently do you use Facebook? != Never]
How likely would you be to "like" or "share" this article in your Facebook News Feed?
-Not at all likely (1)
-Not very likely (2)
-Somewhat likely (3)
-Very likely (4)

We also include questions about topical misperceptions to test our observational hypotheses about prior exposure to misinformation or fake news. We have coded partisan favorability of these statements. We include one false and one true statement favorable to each party. The wording and coding appears below.

To the best of your knowledge, how accurate are the following statements? Each one concerns the allegations of sexual assault made by Christine Blasey Ford against Brett Kavanaugh, President Trump's nominee to the Supreme Court, in a Senate hearing.

The audience at a public rally laughed when Trump mocked gaps in Ford's testimony. (true, pro-R)
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

Ford's allegations were refuted by the people she says were present during the assault. (false, pro-R)
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

Ford's high school classmates recall hearing the story about the alleged assault at the time. (false, pro-D)
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

Kavanaugh was questioned by police after a bar fight in college. (true, pro-D)
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

Wave 2 (survey)

Wave 2 also includes a number of primary outcome measures.

First, we measure beliefs in the claims promoted in the two fake news stimulus articles:

The international financier and philanthropist George Soros has helped to support the caravan of more than 7,000 Central American migrants that is currently moving through Mexico toward the U.S. border.
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

The Trump administration helped Saudi Arabia to target Jamal Khashoggi, the writer for The Washington Post who was recently killed by Saudi agents.
-Not at all accurate (1)
-Not very accurate (2)
-Somewhat accurate (3)
-Very accurate (4)

Second, we measure affective polarization:

Feeling thermometers:
-Democratic Party (0-100)
-Republican Party (0-100)

Affective polarization is calculated as the difference between in-party and out-party ratings. When examining affective polarization, in addition to the difference between in- and out-party ratings, we will also model in-party and out-party ratings independently and report the results separately.

Third, we measure feelings toward the media using a feeling thermometer (0-100).

Fourth, we measure affect toward Jews using a feeling thermometer (0-100).

We also measure voting/intent to vote and to take political action. First, we measure respondents' self-reported voting/intention to vote in the 2018 midterm election:

Have you already voted in the upcoming midterm election?
-Yes (5)
-No

(if no)
Do you, yourself, plan to vote in the upcoming midterm election, or not?
-Yes
-No (1)
-Don't know (2)

(if yes)
How certain are you that you will vote?
-Absolutely certain (4)
-Fairly certain (3)
-Not certain (2)

Due to uncertainty about the proper coding of respondents who are not registered or who answer don't know, we will also analyze a binary measure of intent to vote where 1=already voted/absolutely certain and 0=all other responses.

Intent to take political action is measured as the mean of the following five ANES items:
-I would be willing to place a bumper sticker on my car or wear a campaign button
-I would be willing to volunteer to work for a political campaign
-I would be willing to attend a political rally
-I would be willing to talk to other people about how they should vote
-I would be willing to donate money to a political campaign
-Agree strongly (5)
-Agree somewhat (4)
-Neither agree nor disagree (3)
-Disagree somewhat (2)
-Disagree strongly (1)

<u>Pulse data</u>

We will code respondents' Pulse data for the seven days after they complete the Wave 1 survey as follows:

- Mainstream news visit: One of AOL, ABC News, CBSNews.com, CNN.com, FiveThirtyEight, FoxNews.com, Huffington Post, MSN.com, NBCNews.com, NYTimes.com, Politico, RealClearPolitics, Talking Points Memo, The Weekly Standard, WashingtonPost.com, WSJ.com, or Wikipedia

- Fact-checking visit: PolitiFact, Snopes, or Factcheck.org; the Washington Post Fact Checker is excluded because it is part of the Washington Post, which is already a qualifying media outlet per above
- Fake news visit: Any visit to one of the 673 domains identified in Allcott, Gentzkow, and Yu 2018 as a fake news producer as of September 2018 excluding those with print versions (including but not limited to Express, the British tabloid) and also domains that were previously classified by Bakshy et al. (2015) as a source of hard news.

We will compute a binary measure of exposure to the types of content above as well as a count of the total webpages visited from each category during the period. We will code fake news visits as pro-Democrat (pro-Republican) if 60% or more of the visits by partisans (not including leaners) in our sample period are from Democrats (Republicans). We may also estimate the share of people's news diet from fake news websites as a share of hard news websites and fake news websites visited (using the definition above for fake news and the Bakshy et al. hard news topics classification). Duplicate visits to webpages will not be counted if they are successive (i.e., a page that was reloaded after first opening it). URLs are cleaned of referrer information and other parameters before de-duplication. (For more detail, see the processing steps described in Guess, Nyhan, and Reifler N.d.) Finally, per Allcott, Gentzkow, and Yu 2018, we will include robustness tests in our appendix that replicate each of the hypothesis tests described below using only sites that appear on two or more of their source lists (116 of 673).

Independent variables

Congeniality: We coded news content as pro-Republican or pro-Democrat; we cross this coding with a measure of respondent partisanship to determine if content is congenial (1) or uncongenial (0). (Models including measures of whether content is congenial will omit pure independents for whom this variable is undefined. To provide a strict comparison, we will sometimes estimate main effects models for the set of partisans included in these models.)

Partisanship: We will create two dichotomous independent variables (0/1) for those who identify with the Democratic and Republican parties including leaners (based on ANES party ID questions from the YouGov panel).

Cognitive reflection/analytical thinking and reasoning will be measured by taking the total score on the three-item standard Cognitive Reflection Test (CRT) (Frederick, 2005), using alternate wording (Patel, 2017), and the four-item non-numeric CRT-2 (Thomson, 2016). Both will be presented in multiple choice format (Patel, 2017) with the intuitive but incorrect choice listed first. Correct responses will be scored as 1 and all other responses as 0 and the answer summed. Reliability will be calculated for the combined scale.

Political interest:
-Most of the time (4)
-Some of the time (3)
-Only now and then (2)
-Hardly at all (1)

Political knowledge: Using the survey questions attached below, we will create a scale measuring political knowledge that ranges from 0 (no questions correct) to 8 (all questions correct).

Democrat (0/1): if they indicated they identify as a Democrat or lean toward the Democratic Party

Republican (0/1): if they indicated they identify as a Republican or lean toward the Republican Party

Feelings toward Trump and the media (0-100)

Media trust and confidence:
-A great deal (4)
-A fair amount (3)
-Not very much (2)
-None at all (1)

Conspiracy predispositions - mean of four items:
-Much of our lives are being controlled by plots hatched in secret places.
-Even though we live in a democracy, a few people will always run things anyway.
-The people who really 'run' the country are not known to the voter.
-Big events like wars, recessions, and the outcomes of elections are controlled by small groups of people who are working in secret against the rest of us.
-Strongly agree (5)
-Somewhat agree (4)
-Neither disagree nor disagree (3)
-Somewhat disagree (2)
-Strongly disagree (1)

Reliability will be calculated for the combined scale.

Fake news exposure: We will measure binary and total fake news exposure using the definition above both overall and by whether it is pro- or counter-attitudinal (which we will code using respondent's partisan identity and the coding of the site described above). (If skew is too extreme for the count measure of total fake news exposure, we may use percentage of the news diet instead.) A measure of total fake news exposure prior to wave 1 will also be included as a covariate in tests of H-A4a - H-A4d to make the selection on observables assumption more plausible. (Any exposure to topical misinformation on fake news sites will be excluded from this total.)

Selective exposure tendencies: We will follow Guess et al. in dividing the sample by decile based on the overall average slant of the webpages they visit. (See article for details.) We will also include decile indicators for the period prior to the wave 1 survey as covariates in our observational analyses. (We will confirm that the results from the decile indicator controls are robust to an alternative estimation strategy proposed by Hainmueller, Mummolo, and Xu of using kernel regression to estimate how the marginal effect of the average slant measure varies over its range [our decile indicators are a version of the binning strategy they also recommend].)

Finally, we will also include controls in our observational models for Democrats and Republicans (including leaners), political knowledge (0-8) and interest (1-4), having a four-year college degree (0/1), self-identifying as a female (0/1) or non-white (0/1), and age group dummies (30-44, 45-59, 60+, 18-29 omitted).

Statistical analyses

All headline-level results will be estimated as pooled models using OLS with robust standard errors clustered by respondent and will include question fixed effects (to control for the overall level of credibility of each article) in addition to the covariates specified below. These results will be verified for robustness using appropriate GLM estimators (see below). Models that include indicators for congenial beliefs will omit question fixed effects due to collinearity. For outcomes that are measured at the respondent level rather than the respondent-question level, question fixed effects and clustering will be omitted. For each hypothesis or RQ below describing a series of identical models for real, fake, and hyperpartisan news, we will be estimate the models separately for expositional reasons but will also estimate a pooled model with interactions to determine if there are differences in treatment effects by news type. (These models will include news type dummies and appropriate interactions rather than question fixed effects.)

<u>Observational hypotheses</u>

For H-A1, the outcome measure is exposure to fake news (binary/count/share of information diet):

Fake news exposure = [constant] + selective exposure decile indicators + covariates listed above

For H-A2a, H-A2b, and H-A2c, the outcome measure is the perceived accuracy of fake headlines:

H-A2a: Fake news accuracy = [constant] + prior fake news exposure + covariates listed above

H-A2b: Fake news accuracy = [constant] + prior fake news exposure + congenial + prior fake + news exposure * congenial + covariates listed above

H-A2c: Fake news accuracy = [constant] + prior fake news exposure + CRT score + prior fake + news exposure * CRT score + covariates listed above

For H-A2d, the outcome measure = (mean perceived accuracy of real news headlines - mean perceived accuracy of fake news headlines). This hypothesis is measured at the respondent level using a single mean difference that is not ordered and thus we will only test it using OLS with robust standard errors (i.e., no question fixed effects or clustering).

For H-A3a, H-A3b, and H-A3c, the outcome measure is the perceived accuracy of true and false topical statements. These models will be estimated separately for wave 1 and wave 2 topical misperceptions. For each of these types of statements in wave 1, we will estimate the following models:

H-A3a: Accuracy = [constant] + prior fake news exposure + covariates listed above

H-A3b: Accuracy = [constant] + prior fake news exposure + congenial + prior fake news exposure * congenial + covariates listed above

H-A3c: Accuracy = [constant] + prior fake news exposure + CRT score + prior fake news exposure * CRT score + covariates listed above

For H-A3d, the outcome measure = (mean perceived accuracy of true statements - mean perceived accuracy of false statements). This hypothesis is measured at the respondent level using a single mean difference that is not ordered and thus we will only test it using OLS with robust standard errors (i.e., no question fixed effects or clustering).

Outcome = [constant] + prior fake news exposure + covariates listed above

For wave 2, we will control for treatment assignment in the models described above:

H-A3a: Accuracy = [constant] + prior fake news exposure + congenial fake news exposure + uncongenial fake news exposure + covariates listed above

H-A3b: Accuracy = [constant] + prior fake news exposure + congenial + prior fake news exposure * congenial + congenial fake news exposure + uncongenial fake news exposure + covariates listed above

H-A3c: Accuracy = [constant] + prior fake news exposure + CRT score + prior fake news exposure * CRT score + congenial fake news exposure + uncongenial fake news exposure + covariates listed above

H-A3d: Outcome = [constant] + prior fake news exposure + congenial fake news exposure + uncongenial fake news exposure + covariates listed above

~~Main effects of news tips intervention~~

~~For H-B1a, the outcome measure is the perceived accuracy of fake headlines.~~

~~For H-B1b, the outcome measure is the perceived accuracy of real headlines.~~

~~For H-B1d, the outcome measure is the perceived accuracy of hyper-partisan headlines.~~

~~For RQ-B1, the outcome measures are binary measures and counts of visits to fake news websites (RQ-B1a), fact-checking websites (RQ-B1b), and mainstream news websites (RQ-B1c).~~

~~For RQ-B2, the outcome measures are self-reported intention to share real, fake, and hyper-partisan news.~~

~~For each of these question-level hypotheses, we will estimate the following model using OLS regression (with robustness checks using ordered probit or count models as appropriate):~~

~~Outcome = [constant] + News tips~~

~~For H-B1c, the outcome measure = (mean perceived accuracy of real news - mean perceived accuracy of fake news). This hypothesis is measured at the respondent level using a single mean difference that is not ordered and thus we will only test it using OLS with robust standard errors (i.e., no question fixed effects or clustering).~~

~~Outcome = [constant] + News tips~~

~~Heterogeneous treatment effects~~

~~For H-C1, the outcome measure is the perceived accuracy of fake, real, and hyper-partisan headlines. For each of these types of headlines, we will estimate the following model:~~

~~Outcome = [constant] + News tips + Congenial + News tips*congeniality~~

~~We will also explore question-level results to test for asymmetry. So for each question we will estimate the model:~~

~~Outcome = [constant] + News tips + Party FE + News tips*party FE~~

~~For the exploratory analyses of possible moderators (Cognitive Reflection Test scores (per Pennycook and Rand 2018), trust in and feelings toward the media, feelings toward Trump, conspiracy predispositions, political interest and knowledge, and pre-treatment visits to fake news sites and fact-checking sites), the outcome measure is the perceived accuracy of fake, real, and hyper-partisan headlines. Our focus will be on whether the results from our analysis of H-B1a, H-B1b, and H-B1d above are moderated by these individual differences. That is, we want to see if the effect is moderated when the outcome is the (1) perceived accuracy of fake headlines, (2) the perceived accuracy of the real headlines, and (3) the perceived accuracy of hyper-partisan headlines. Thus we will be looking at how each of the potential moderators interacts with the treatments for all three of the outcomes specified above.~~

~~Due to likely collinearity between the predictors, we will estimate separate models of perceived accuracy for each potential moderator.~~

~~e.g.:~~

~~Outcome = [constant] + News tips + Political interest + News tips*political interest~~

~~Outcome = [constant] + News tips + Congenial + Political interest + News tips*political interest + Congenial*political interest + News tips*congenial + News tips*congenial*political interest~~

Waves 1-2

For H-D1, RQ-D1, and RQ-D2, the outcome measure is the perceived accuracy of fake headlines in Wave 2.

For H-D2 and RQ-D3, the outcome measure is the perceived accuracy of real headlines

For H-D3 and RQ-D4, the outcome measure is the perceived accuracy of hyper-partisan headlines.

To test the effects of wave 1 exposure (H-D1-H-D3)

Outcome = [constant] + wave 1 exposure

~~For RQ-D1, the the outcome measure is the perceived accuracy of fake headlines in Wave 2, but we will test if the treatment is moderated by exposure to the news tips in Wave 1.~~

~~Outcome = [constant] + wave 1 exposure + News tips + wave 1*News tips~~

~~For the question-level hypotheses RQ-D2, RQ-D3, and RQ-D5, we will estimate the following model using OLS regression (with robustness checks using ordered probit as appropriate):~~

~~Outcome = [constant] + News tips~~

~~For RQ-D4, the outcome measure = (mean perceived accuracy of real news - mean perceived accuracy of fake news). This hypothesis is measured at the respondent level using a single mean difference that is not ordered and thus we will only test it using OLS with robust standard errors (i.e., no question fixed effects or clustering).~~

~~Outcome = [constant] + News tips~~

<u>Main effects of fake news exposure</u>

For H-E1a and RQ-F1, the outcome measure is affective polarization.

For H-E1b and RQ-F2, the outcome measure is affect toward the media.

For H-E2 and RQ-F3, the outcome measure is intent to vote.

For H-E3 and RQ-F4, the outcome measure is the intent to take political action scale.

For each of these hypotheses, we will estimate the following model using OLS regression (with robustness checks using ordered probit where appropriate):

Outcome = [constant] + congenial fake news exposure + uncongenial fake news exposure

For H-E4, the outcome measures are beliefs in the Soros/caravan and Trump/Khashoggi conspiracy theories. We will test the main effect of the articles on beliefs in these claims using separate OLS regressions for each outcome measure:

Outcome = [constant] + caravan article fake news exposure + Khashoggi article fake news exposure

For RQ-E1a, the outcome measure is feelings toward Jews. We will test the main effect of the article using OLS regression:

Outcome = [constant] + caravan article fake news exposure + Khashoggi article fake news exposure

Heterogeneous treatment effects

For the exploratory analyses of possible moderators of the effects of congenial fake news exposure, the outcome measures are affective polarization, intent to vote, intent to take political action, trust in and feelings toward the media, feelings toward Trump (entered as a linear term and with indicators for terciles or quartiles), conspiracy predispositions, political interest and knowledge, pre-treatment visits to fake news sites and fact-checking sites, and Hispanic/Latino-white and Christian-Muslim differences in feeling thermometer scores. Due to likely collinearity between the predictors, we will estimate separate models for each potential moderator for each outcome measure.

E.g.:

Outcome = [constant] + congenial fake news exposure + feelings toward Trump + congenial fake news exposure * feelings toward Trump + uncongenial fake news exposure

For H-E5, the outcome measures are beliefs in the Soros/caravan and Trump/Khashoggi conspiracy theories. We will test if the effects of fake news articles on these beliefs vary by whether the article is pro-attitudinal using separate OLS regressions among Democrats or Republicans defined using the coding above:

Outcome = [constant] + caravan article fake news exposure + Trump Khashoggi article fake news exposure + Republican + caravan*Republican + Khashoggi*Republican

For RQ-E1b, we will also test if the effects of caravan article exposure vary by whether the article is pro-attitudinal using OLS regression among Democrats or Republicans defined using the coding above:

Outcome = [constant] + caravan article fake news exposure + Trump Khashoggi article fake news exposure + Republican + caravan*Republican + Khashoggi*Republican

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 (forthcoming) and replace any continuous interactions in our models with them if it does not.

-We will compute and report summary statistics for our sample. 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 or count dependent variables will be verified for robustness using ordered probit or Poisson regression with standard errors, respectively.

-We may estimate the experimental models described above with a standard set of covariates if including those has a substantively important effect on the precision of our treatment effect estimates. We will select covariates from the list below using the lasso before estimating the model using OLS per the procedures described in Bloniarz et al. 2016.

> Candidate covariates for all models:
> -average media slant value from the period prior to wave 1 (measured per Guess et al. N.d.)
> -decile indicators for average media slant from the period prior to wave 1 (measured per Guess et al. N.d.)
> -indicators for Democrats and Republicans (including leaners)
> -gender
> -age groups (30-44, 45-59, 60+)
> -non-white respondents
> -respondents with a four-year college degree
> -scores on standard political knowledge and interest scales
> -Trump feeling thermometer from wave 1
> -Media feeling thermometer from wave 1
> -Trust in the media in wave 1
> -Affective polarization in wave 1 (in-party feelings - out-party feelings)
> -Conspiracy predispositions scale (average response in wave 1)
>
> Added covariate for wave 2 belief accuracy outcome models:
> -lagged average belief accuracy from wave 1 for each type of outcome measure (e.g., average accuracy of pro-Republican fake news in wave 1 for analysis of perceived

accuracy of pro-Republican fake news in wave 2; applies to all real/fake/hyperpartisan news types)

-We will test for differential attrition between survey waves by examining the relationship between completion of wave 2 and wave 1 treatment assignment. This is our primary measure of attrition. If we observe significant differential attrition based on condition, we will use a strategy such as the one proposed by Aronow et al. to account for missing outcome variables in a randomized experiment (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2305788).
We will also test for attrition based on the following observable characteristics:
-media trust, media feelings, and average fake news belief in wave 1
-political characteristics (partisanship and political knowledge)
-demographic characteristics (race, sex, and age)

Due to the large number of characteristics on which we will assess imbalance, we will use a correction for multiple comparisons. Attrition that is unrelated to random assignment should not confound our treatment effect estimates but is worth noting for the reader and potentially addressing in any observational data analyses.

-In addition to OLS with robust standard errors and question fixed effects, we will also fit a non-nested hierarchical model with the same covariates.

-We will also compute and report descriptive statistics for our data to summarize sample characteristics, response variable distributions, etc.

Country

United States

Sample Size (# of Units)

2857.

Was a power analysis conducted prior to data collection?

No

Has this research received Institutional Review Board (IRB) or ethics committee approval?

Yes

IRB Number – Michigan HUM00153414, WUSTL 201806142 (amended), Princeton 10875-02,, no approval number at Exeter (accepts IRBs from elsewhere)

Date of IRB Approval – October 19, 2018 (Michigan), October 10, 2018 (WUSTL), October 9, 2018 (Princeton)

Will the intervention be implemented by the researcher or a third party? If a third party, please provide the name.

Other: YouGov

Did any of the research team receive remuneration from the implementing agency for taking part in this research?

Yes

If relevant, is there an advance agreement with the implementation group that all results can be published?

Yes (informal)

JEL classification(s)

N/A