Political Behavior
https://doi.org/10.1007/s11109-019-09533-0

ORIGINAL PAPER

Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media

Katherine Clayton[1] · Spencer Blair[1] · Jonathan A. Busam[1] · Samuel Forstner[1] · John Glance[1] · Guy Green[1] · Anna Kawata[1] · Akhila Kovvuri[1] · Jonathan Martin[1] · Evan Morgan[1] · Morgan Sandhu[1] · Rachel Sang[1] · Rachel Scholz-Bright[1] · Austin T. Welch[1] · Andrew G. Wolff[1] · Amanda Zhou[1] · Brendan Nyhan[2]

Abstract

Social media has increasingly enabled "fake news" to circulate widely, most notably during the 2016 U.S. presidential campaign. These intentionally false or misleading stories threaten the democratic goal of a well-informed electorate. This study evaluates the effectiveness of strategies that could be used by Facebook and other social media to counter false stories. Results from a pre-registered experiment indicate that false headlines are perceived as less accurate when people receive a general warning about misleading information on social media or when specific headlines are accompanied by a "Disputed" or "Rated false" tag. Though the magnitudes of these effects are relatively modest, they generally do not vary by whether headlines were congenial to respondents' political views. In addition, we find that adding a "Rated false" tag to an article headline lowers its perceived accuracy more than adding a "Disputed" tag (Facebook's original approach) relative to a control condition. Finally, though exposure to the "Disputed" or "Rated false" tags did not affect the perceived accuracy of unlabeled false or true headlines, exposure to a general warning *decreased* belief in the accuracy of true headlines, suggesting the need for further research into how to most effectively counter false news without distorting belief in true information.

Keywords Fake news · Fact check · Warnings · Corrections · Social media · Misperceptions

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11109-019-09533-0) contains supplementary material, which is available to authorized users.

✉ Brendan Nyhan
bnyhan@umich.edu

Extended author information available on the last page of the article

Published online: 11 February 2019

Since the 2016 U.S. election, the effects of "fake news" have received considerable attention. Many Americans now worry about the effects of this factually dubious content that imitates the format of journalism but is produced with no regard for accuracy or fairness (Lazer et al. 2018). This type of content, which we will refer to as "false news" for expositional clarity, is most often created for profit by dubious websites.[1] Many people appear to believe false news stories, which circulated widely before the election and generally favored Donald Trump over Hillary Clinton (Silverman 2016; Silverman and Singer-Vine 2016). Although false news most likely did not change the election's outcome (Allcott and Gentzkow 2017), its prevalence is still an important concern. False news promotes misperceptions among voters and can induce distrust of legitimate information. In this sense, it presents a serious threat to American democracy.

The public's vulnerability to false information has grown as people have come to increasingly rely on social media as a source of news. According to a recent Pew survey, 62% of American adults get news from social media sites such as Facebook (Gottfried and Shearer 2017), which played an especially important role in the spread of false news during the 2016 presidential campaign. The most viral false news articles were shared more on Facebook in the months prior to the election than the most widely shared mainstream news stories (Silverman 2016). Online misinformation, both political and otherwise, has continued to be a challenge since the election. For example, false claims swirled around social media in the aftermath of Hurricane Harvey, including an article claiming that Black Lives Matter protesters blocked emergency responders from reaching hurricane victims (Schaedel 2017).

To address these concerns, Facebook began adding "Disputed" tags to stories in its News Feed that have been debunked by fact-checkers in December 2016 (Mosseri 2016). It used this approach for approximately one year before switching to providing fact-checks in a "Related Articles" format underneath suspect stories (Smith et al. 2017). The company also promoted tips for spotting false news at the top of News Feed in April 2017 and May 2018 (Constine 2017; Owen 2018). Both approaches presumably seek to reduce the probability that people will believe false news articles.

Research suggests that combating misinformation is a difficult challenge (for reviews, see, e.g., Flynn et al. 2017; Lewandowsky et al. 2012). In particular, studies that focus specifically on exposure to false news on social media have found mixed results. Though "disputed" tags seem to modestly reduce belief in false news headlines, they may fail to counteract exposure effects over time (Pennycook et al. 2017) and could create an "implied truth" effect in which unlabeled false headlines are seen as more accurate (Pennycook and Rand 2017). Similarly, Ecker et al. (2010)

[1] "Fake news" has many definitions and is frequently used in imprecise or confusing ways. Moreover, the debate over the meaning of the term and related concepts raises epistemological issues that are beyond the scope of this paper (e.g., speaker intent; see Wardle and Derakhshan 2017). We therefore employ "false news" as an alternative term throughout this paper, which define as described above ("factually dubious content that imitates the format of journalism but is produced with no regard for accuracy or fairness"; see Lazer et al. 2018). This approach is consistent with the practices of various news and social media sources (e.g., Oremus 2017) and is intended to avoid unnecessary confusion.

find that specific warnings are more effective than general warnings at reducing the continued influence of exposure to false information on beliefs, but neither approach eliminates this effect entirely. They argue that a specific warning (directly alerting readers about how misinformation can continue to influence them even after being debunked) reduces belief in false claims by helping people to tag misinformation, whereas a general warning (telling participants that the media sometimes does not check facts before publishing information that turns out to be inaccurate) promotes "nonspecifically induced alertness" (Ecker et al. 2010, p. 1096) that is less effective.

In this study, we investigate whether interventions like the ones used by Facebook can effectively reduce belief in false news. Specifically, we test the effects of both a general warning about false news and two types of specific warnings about individual articles questioned by fact-checkers. Our results indicate that exposure to a general warning about false news modestly reduces the perceived accuracy of false headlines. We also find that adding a "Rated false" or "Disputed" tag underneath headlines reduces their perceived accuracy somewhat more. In particular, the "Rated false" tag is most effective at reducing the perceived accuracy of false headlines, though neither tag measurably reduced the self-reported likelihood that headlines would be shared on social media. The effects of these tags did not vary consistently depending on whether participants had previously received a general warning. Similarly, there were not consistent differences between the effect of tags on politically congenial versus non-congenial information. Finally, though we find no evidence that tagging headlines as "Rated false" or "Disputed" has large spillover effects to belief in other headlines, exposure to a general warning did reduce belief in the accuracy of true headlines as well as false ones, suggesting that efforts to promote greater skepticism toward false news can also increase distrust of legitimate news and information.

Theoretical Expectations

We specifically test the following hypotheses and research questions, which were pre-registered at EGAP prior to the administration of our study (http://www.egap.org/registration/2516).

First, though people's initial belief in false information can be difficult to change (see Flynn et al. 2017 for a review), some evidence suggests that warnings about false information can reduce belief in false claims or prevent the uptake of misinformation. Ecker et al. (2010) find that warnings about the limits of fact-checking in the media reduce belief in outdated facts and increased acceptance of correct information, but do not entirely eliminate the effect of misinformation. Similarly, Bolsen and Druckman (2015) find that warnings are more effective than corrections at countering directionally motivated reasoning about scientific claims.

We focus specifically on headlines, which are the dominant form of content in social media and can be misleading even to relatively attentive readers (e.g., Ecker et al. 2014). Our study tests the effectiveness of two approaches that have been used by Facebook to try to reduce belief in false news: general warnings to beware of misleading content and specific tags on article headlines that mark

them as "Disputed." We also test the effectiveness of specific tags that instead mark headlines as "Rated false."

The first approach we test is a general warning. In April 2017 and May 2018, Facebook rolled out a warning of this sort to users, distributing a message at the top of News Feed that highlighted "tips for spotting fake news" (Mosseri 2017; Owen 2018). In this experiment, we use an analogous warning message to test the following hypothesis:

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

Our study also tests the effect of a specific warning by building on Pennycook et al. (2017), who find that a Facebook-style "Disputed" tag under headlines reduces belief in the accuracy of false stories and reduces users' intent to share them. Similarly, Bode and Vraga (2015) find that including corrective information in Facebook's "related stories" function, which links articles to other articles that may correct false claims, effectively reduces misperceptions. These interventions warn users about misinformation or false news at the time when they are exposed to a headline and are intended to help readers notice false information as soon as they encounter it. We therefore propose the following hypothesis:

> H2a: The presence of a Facebook-style "Disputed" tag under false headlines will reduce their perceived accuracy relative to a no-tag condition.

However, tags warning that a claim is "Disputed" may not be sufficiently direct. We therefore evaluate the effect of a specific warning directly stating that a false news headline is untrue in an additional condition. A warning of this nature, though not yet used by Facebook, might convey a stronger message than the inconclusive terminology of the "Disputed" warning. Communicating expert consensus has been found to increase belief in global warming and support for expert views on other environmental problems (Aklin and Urpelainen 2014; Bolsen and Druckman 2015; Corbett and Durfee 2004). Personal agreement with the existence of global warming rises consistently with communicated levels of scientific agreement (Chinn et al. 2018a). Moreover, media coverage that does not take a side in an effort to appear "balanced" can distort public perceptions of expert consensus (Boykoff and Boykoff 2004; Koehler 2016). Finally, audiences make inferences about the communicators of information and the context that information comes from when they process new information (e.g., Wegner et al. 1981). In the context of false news on Facebook, "Disputed" tags may signal ambivalence about how strongly the platform endorses the fact-check.

Our alternate specific warning describes a false news headline as "Rated False by Snopes and Politifact." This tag is specific enough to effectively reduce belief (Ecker et al. 2010) and more clearly conveys the consensus among fact-checking websites that the claim in the article is false. We therefore expect that the effects of the "Rated false" tag would be larger than the effects of the "Disputed" tag and propose the following hypothesis:

H2b: The presence of a "Rated false" tag under false headlines will reduce their perceived accuracy relative to a no-tag condition.
H2c: The presence of a "Rated false" tag under false headlines will reduce their perceived accuracy relative to a Facebook-style "Disputed" tag.

Ecker et al. (2010) do not test how general and specific warnings work in tandem with one another, though it is plausible that a general warning could increase alertness to subsequent specific warnings about false information. Indeed, van der Linden et al. (2017) find that presenting respondents with information on the scientific consensus about global warming, as well as a general or specific statement about the existence of climate change, was more effective at inoculating respondents against misinformation on climate change than either treatment alone. We therefore propose the following hypothesis about how exposure to a general warning will strengthen the effects of specific warnings:

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

Consistent with prior research (Flynn et al. 2017; Kahan 2015), we also expect that people's belief in false news depends on whether it aligns with their political identity and preferences. "Disputed" or "Rated false" tags could be less effective when a person is viewing a headline with which they are inclined to agree (e.g., Nyhan and Reifler 2010; Kahan et al. 2017). In this case, we evaluate how the effects of our experimental manipulations vary depending on participants' approval of President Trump (each article concerns either President Trump and his allies or his opponent, Hillary Clinton) and the slant of the articles in question. For example, the negative effect of a "Disputed" or "Rated false" tag on the perceived accuracy of a news story may be attenuated if the news story is politically congenial (e.g., a pro-Trump headline seen by a Trump supporter). We therefore hypothesize that the effect of specific warnings will be reduced when they accompany politically congenial information relative to uncongenial information:

H4: The effect of a Facebook-style "Disputed" tag on perceived accuracy (H4a) or "Rated false" tag (H4b) will be reduced for politically congenial information versus uncongenial information (versus a headline with no tag).

Finally, we also seek to answer three pre-registered research questions about which we had weaker theoretical expectations. Drawing on previous research identifying the importance of political preferences on belief in misinformation, we investigate whether the effect of warnings on the perceived accuracy of headlines varies between congenial information and uncongenial information (RQ1). We also investigate whether specific warnings on a headline will affect the perceived accuracy of untagged false (RQ2a)[2] or true (RQ2b) headlines, and whether a general warning

[2] Pennycook and Rand (2017), which we had not seen at the time of pre-registration, also considers this question.

about misleading articles will reduce the perceived accuracy of true information (RQ3). Practically, it would be difficult for social media platforms to fact-check and add a "Disputed" or "Rated false" tag to every false news headline. Because some false news headlines could inevitably fall through the cracks, we are interested in seeing how general and specific warnings influence respondents' perceived accuracy of such news items.[3]

Methods

Participants

The study, which was approved by the Dartmouth College Committee for the Protection of Human Subjects (STUDY0003028), was conducted from May 8–9, 2017 among participants recruited from Amazon Mechanical Turk (MTurk). Although samples from MTurk are not nationally representative, results from studies conducted with participants from the site mirror those obtained from other samples (e.g., Berinsky et al. 2012; Coppock 2016; Horton et al. 2011; Mullinix et al. 2015).[4] Non-U.S. residents, people under 18 years of age, and people who completed a prior pilot study were not allowed to participate.[5] We also exclude six respondents from the data who dropped out prior to the experimental manipulation. Our final sample is 2994 participants.

Although our sample is diverse, it skews female (54% female), younger (median age group 25–34) and more educated (55% have a bachelor's degree or greater) than the U.S. population. Our sample also overrepresents Democrats—32% identify as Republican or lean Republican, whereas 58% identified as Democrat or lean Democrat. Participants also approve of Trump (30%) and voted for him (30% of those who report voting) at lower rates than the U.S. population. A detailed comparison of the composition of our sample to population benchmarks is provided in Online Appendix C (Table C1).

[3] We pre-registered an additional research question about the effects of exposure to a general warning and/or to a "Disputed" or "Rated false" tag on respondents' self-reported likelihood of "liking" and sharing the headlines on Facebook. The results of this analysis are presented in Online Appendix B.

[4] A minority of studies conclude that MTurk samples are not externally valid (e.g., Krupnikov and Levine 2014). For example, participants on MTurk tend to skew liberal and young. Moreover, the underrepresentation of conservatives and older participants may suggest that these participants differ from other conservatives or older individuals in the general population. However, numerous studies find that experimental treatment effect estimates typically generalize from MTurk to national probability samples, suggesting these problems are rare (e.g., Berinsky et al. 2012; Coppock 2016; Horton et al. 2011; Mullinix et al. 2015). Finally, our MTurk sample is externally valid in the sense that it is made up disproportionately of frequent users of the Internet—precisely the group who may be most likely to encounter false news (Pennycook and Rand 2018a). We thus conclude that respondents from MTurk constitute a valid sample for testing our hypotheses, though replication on representative samples would of course be desirable.

[5] The pilot study tested the effects of "Disputed" and "Rated false" tags only on perceived accuracy and likelihood of liking/sharing for six false news headlines. The results of this study were similar to our main analysis, and are available upon request.

Table 1 Experimental conditions

Tag	General warning	N
None	No	469
None	Yes	424
"Disputed"	No	413
"Disputed"	Yes	429
"Rated false"	No	429
"Rated false"	Yes	397
Pure control		433

Experimental Design and Procedure

We focus on beliefs in headlines because of their primacy on social media (Gabielkov et al. 2016; Manjoo 2013). The initial judgments that people form when reading headlines are also likely to shape their subsequent beliefs and opinions (Thorson 2016).

The experiment used a 2×3 between-subjects design that also includes a pure control group. Participants were randomly assigned with equal probability to a pure control group or to one of six experimental conditions (see Table 1). We manipulated whether participants were exposed to a general warning about misleading articles or not (middle column of Table 1). We also independently randomized non-controls into one of three headline conditions: a condition in which no fact-checking tags were presented (first two rows of Table 1), a specific warning condition that included tags labeling articles as "Disputed" (second two rows of Table 1), and a specific warning condition in which they were instead labeled as "Rated false" (last two rows of Table 1).

The study proceeded as follows. Once participants consented to participate, they answered a series of demographic questions, followed by questions about their use of social media, political preferences, voting behavior, and trust in fact-checking and the media. Participants were then asked to rate the accuracy of several real and fabricated political statements to test their predisposition to hold political misperceptions.[6] Afterward, they answered a political knowledge battery, which provided a buffer between the misperception items and the experimental task.[7]

[6] As in most studies, we cannot know how much false news respondents were exposed to during the 2016 presidential election and its aftermath (e.g., Allcott and Gentzkow 2017). While it would be useful to measure this quantity, our main interest is the effect of warnings and tags on belief accuracy when they encounter false news. In addition, the auxiliary measure of misperception belief mentioned above does allow us to test whether individuals who are susceptible to believing false news respond differently to warnings and tags than those who are not. We find no consistent evidence of such heterogeneity in exploratory analyses reported in Online Appendix C. Scholars should collect data on individuals' exposure to false news and explore treatment effect heterogeneity by this variable directly in future research.

[7] A possible concern is that asking respondents to rate political statements for accuracy could have primed them to be particularly alert to clues that the treatment articles could be deceptive in nature. However, Pennycook et al. (2017) and Pennycook and Rand (2017) did not ask respondents to rate any statements for accuracy before their experiment and also found that tagged false news headlines were

In the general warning condition, participants were shown a message warning them about misleading articles and providing advice for identifying false information (see Online Appendix A for exact wording and design). The design of the general warning was chosen to resemble Facebook's false news message to users (Mosseri 2017). Participants in the no-warning conditions were shown an identical image with innocuous instructions to eliminate any potential confounding effects. In the pure control group, respondents were exposed to no images, no articles, no general warning, no tags, and no headlines, and proceeded directly to the questions measuring the outcome variable (discussed in the next section).

Each participant who was not assigned to the pure control group was shown nine selected political headlines formatted as they would appear on Facebook in random order: three false pro-Trump headlines, three false anti-Trump headlines, and three true headlines (Table 2; see Online Appendix A for the exact stimuli used). Each participant saw each of the nine headlines, with the appearance of the headlines (i.e., whether they included "Disputed" or "Rated false" tags) randomly varying based on treatment condition. We selected a balance of pro- and anti-Trump false news articles from Snopes and Buzzfeed, excluding those related to the 2016 election that had become less relevant.[8] True political headlines from mainstream media sources were also included so that the veracity of headlines was not uniform.[9] Finally, though Pennycook and Rand (2018b) find that news sources do not significantly affect belief in the perceived accuracy of false news headlines (see also Clayton et al. 2018a), we purposefully omitted news sources (and authors) to minimize potentially confounding variables and isolate the effects of warnings and tags on belief in false news headlines.

In the disputed condition, two randomly chosen pro-Trump and two anti-Trump false news headlines were tagged as "Disputed by Snopes.com and PolitiFact" (see Online Appendix A for headline format). Similarly, in the false condition, two randomly chosen pro-Trump and two anti-Trump false news headlines were tagged as "Rated false by Snopes.com and PolitiFact." The wording and format of these tags were chosen to resemble warnings implemented by Facebook. Tags were distributed evenly to pro-Trump and anti-Trump headlines (i.e., two of each). Finally, the two remaining false headlines (one pro-Trump and one anti-Trump) were not tagged. This distribution allows us to test the effects of political congeniality while also simulating a typical news feed in which not all false news stories will be fact-checked.

Footnote 7 (continued)
rated as less accurate that untagged ones, suggesting that the tags reduce the perceived accuracy of false headlines independently of a possible priming effect.

[8] Some of these articles were originally used in Pennycook et al. (2017), which examined the effect of prior exposure to false news headlines on the perceived accuracy of false news. Others were taken from Silverman (2016), a compilation of the most widely shared false news articles during the 2016 election. The original sources of the false news articles were dubious websites that had intentionally created them for profit.

[9] The true headlines that were tested were taken from actual mainstream news sources and were not intended to be explicitly pro- or anti-Trump, though respondent interpretations of them may differ.

Table 2 Headlines displayed in survey

Headline	Source	Type
Trump questions why U.S. Civil War had to happen	Reuters	True
Trump Orders Airstrikes in Syria After Chemical Attack	CBS New York	True
Neil Gorsuch Confirmed to Supreme Court	CNN	True
Trump on Revamping the Military: "We're Bringing Back the Draft"	Real News Right Now	False, anti-Trump
Trump Plagiarized the Bee Movie for Inaugural Speech	Daily Kos	False, anti-Trump
FBI Discovers Kremlin is blackmailing Jason Chaffetz over Donald Trump and Russia	Palmer Report	False, anti-Trump
"Donald Trump Protester Speaks Out: 'I was paid $3,500 to protest Trump's rally'"	ABCnews.com.co	False, pro-Trump
Donald Trump Sent His Own Plane to Transport 200 Stranded Marines	Top Rated Viral	False, pro-Trump
FBI Agent Suspected in Hillary Email Leaks Found Dead in Apparent Murder-Suicide	Alexander Higgins	False, pro-Trump

In the study, the Chaffetz headline identified him as a "Republican Congressman."

After each headline was displayed, participants were asked to evaluate the accuracy of the headline and to self-report how likely they would be to "like" and share the story on Facebook (see Online Appendix A for wording).

Measures

To test the perceived accuracy of the claims in false news headlines, participants were asked to evaluate the accuracy of each claim on a four-point Likert scale from "Not at all accurate" (1) to "Very accurate" (4).[10] This question format is a common approach in recent studies measuring participants' belief in misinformation and false news (e.g., Clayton et al. 2018a; Kuru et al. 2017; Pennycook et al. 2017; Pennycook and Rand 2017, 2018a, b); employing it allows us to directly compare our results with the existing literature. We reasoned that a unipolar four-point scale allowed respondents to express a more nuanced assessment of a statement's accuracy than, for example, a binary true/false question (e.g., by choosing "somewhat accurate" versus "very accurate"). This question format was also appropriate for the pure control group, which did not view a headline but could be asked to assess each claim's general accuracy. Finally, we elected not to include a "don't know" option and instead permitted respondents to skip questions in the survey.[11] Summary measures of respondent belief in each of the nine false news headlines included in our survey are provided in Online Appendix C (see Table C2).[12]

We also asked participants who indicated they use Facebook in a pre-treatment measure about their willingness to "like" or share a given headline on Facebook on a scale from "Not at all likely" (1) to "Very likely" (4) (see Online Appendix B for results). Finally, because our headlines included a mixture of pro- and anti- Trump stories, we measured respondents' approval of President Trump prior to the manipulation on a scale from "Strongly disapprove" (1) to "Strongly approve" (4) and classified those who "strongly" or "somewhat" approve of him as approvers and those who "strongly" or "somewhat" disapprove as disapprovers.

[10] A potential concern is that highly attentive MTurk respondents saw these accuracy questions as an attention check rather than a measure of sincere belief and responded accordingly. However, previous research has found that the effect of corrections to misinformation were almost identical among samples of MTurk workers and Morning Consult poll respondents (Nyhan et al. 2017) and provides limited and inconsistent evidence of demand effects in survey experiments (Mummolo and Peterson 2018).

[11] de Leeuw et al. (2015) find that excluding "don't know" options but allowing respondents to skip questions in online surveys (as we did) reduces missing data and increases reliability in online surveys relative to the inclusion of a "don't know" option, and suggest using "don't know" options only when there is a theoretical reason to do so. We also opt to exclude the "don't know" option to encourage compatibility between our study and others in the field that examine belief in false news and other forms of political misinformation (e.g., Pennycook et al. 2017; Pennycook and Rand 2017).

[12] Our preregistration did not offer hypotheses about the correlates of false news belief, but see Pennycook and Rand (2018b), which finds that individuals who have a tendency to ascribe profundity to randomly generated sentences and who overstate their level of knowledge are more likely to perceive false news as accurate. Those who engage in analytic thinking are less susceptible.

Results

We analyze the effects of our experiment using OLS with robust standard errors.[13] All analyses were pre-registered in advance on EGAP unless otherwise specified. Replication data and code are available on the *Political Behavior* Dataverse (https://dataverse.harvard.edu/dataverse/polbehavior).

Our primary outcome measure is the perceived accuracy of false headlines, which we pooled across the false news headlines that respondents evaluated. The statistical analyses below include standard errors clustered by respondent and question fixed effects as well as indicators for exposure to a general warning about false news and whether the respondent saw a "Disputed" or "Rated false" tag under the headline in question. Importantly, our pre-registered specifications exclude responses to untagged headlines by respondents in the disputed or false conditions. We thus compare responses to tagged headlines in the disputed and false conditions to responses to headlines in the condition in which no tags were shown. All experimental treatment effects were estimated as intent to treat effects.[14]

Since we could not estimate respondents' likelihood to "like" or share headlines (results in Online Appendix B) in our pure control condition, we deviate from our pre-registration for expositional reasons to present estimates below that exclude the pure control group. As a result, we focus on the sample of 2561 respondents assigned to our experimental conditions. Accordingly, the baseline in our statistical models is the group that did not receive a general warning about false news or any "Disputed" or "Rated false" tags on headlines. Our estimates are substantively identical when the pure control group is used as the baseline condition instead (see Online Appendix C).

Effects of General and Specific Warnings

Figure 1 summarizes the mean perceived accuracy of the false news headlines by whether respondents received a general warning about misleading articles and/or whether the headline was identified as "Disputed" or "Rated false" by fact-checkers. As the figure indicates, a general warning only slightly decreased the perceived accuracy of false headlines in the untagged headlines condition, reducing it from 1.96 to 1.90 on our four-point Likert scale.[15] The perceived accuracy of false headlines declined more when specific warnings were provided, decreasing from 1.96 to 1.73 when a "Disputed" tag appeared and 1.62 when a "Rated false" tag appeared. This effect was nearly identical when a general warning was previously provided. In

[13] All results are virtually identical when estimated using ordered probit instead. See Online Appendix C.

[14] We do not include respondent fixed effects, which were incorrectly specified in the pre-registration (they cannot be estimated due to multicollinearity). However, we show in Online Appendix C that our primary results are consistent when estimated in a model that includes random effects by respondent.

[15] The estimates reported here refer to the effects of each treatment alone independent of any moderators, with all other manipulations set at 0. We estimate models that include interactive terms below.

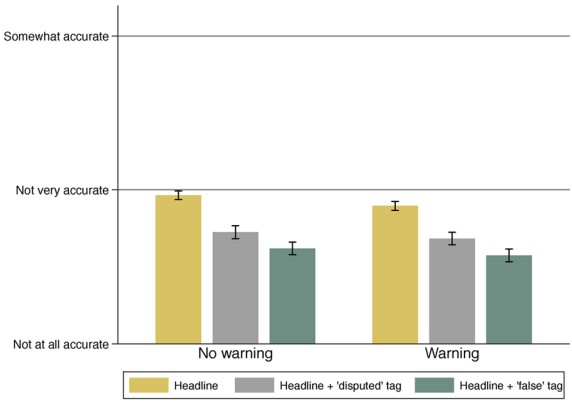

Fig. 1 Effects of general and specific warnings on the perceived accuracy of false headlines. Mean belief that false headlines were accurate on a four-point Likert scale from "Not at all accurate" (1) to "Very accurate" (4). See Online Appendix A for question wording and stimulus materials

substantive terms, the "Disputed" tag reduced the mean proportion of respondents who accept a headline as "Somewhat accurate" or "Very accurate" when no general warning was provided from 29% in the baseline condition to 19%, a ten-percentage point decline (95% CI in a regression using the binary accuracy measure described above as the outcome: −7 to −13 percentage points). Similarly, the "Rated false" tag reduced the proportion of respondents who accepted the headline as accurate to 16%, a 13-percentage point decline from the baseline condition (95% CI −11 to −17 percentage points). These effects are larger than those reported in Pennycook and Rand (2017), who find that a "Disputed" tag reduces perceived accuracy by 3.7 percentage points (a point estimate that is outside our 95% confidence interval).

We test our hypotheses and research questions more formally in Table 3, which shows the results of our pooled regression models for the perceived accuracy of false news headlines. Our results largely support the hypotheses that warnings reduce belief in false information. Consistent with H1, average belief in false headlines was slightly lower for participants who saw a general warning before seeing headlines than for participants who saw headlines with no warning (− 0.08; $p < .05$). However, the substantive magnitude of this reduction in perceived belief accuracy is small (Cohen's d = 0.08).[16]

The negative effect of tags on perceived accuracy was stronger, however. Average perceived accuracy for participants who saw a headline with a "Disputed" tag was 0.24 points lower on our four-point scale than for participants who saw no tag ($p < 0.01$; Cohen's d = 0.26) and 0.34 points lower for those who saw a "Rated false" tag than for those who saw no tag ($p < 0.01$; Cohen's d = 0.38), supporting H2a and H2b. Most notably, "Rated false" tags were significantly more effective than "Disputed" tags at reducing belief in false information relative to a no tag condition (− 0.11, $p < 0.01$), supporting H2c and suggesting that the effect of specific warnings is greater when they clearly indicate that a headline is false.

[16] The effects on perceived accuracy reported in Tables 3–5 are consistent when non-Facebook users are excluded from the sample in exploratory analyses (see Online Appendix C).

Table 3 Experimental effects on perceived accuracy of false headlines

	Accuracy
General warning	−0.08**
	(0.03)
"Disputed" tag	−0.24***
	(0.04)
"Disputed" × warning	0.04
	(0.05)
"Rated false" tag	−0.34***
	(0.04)
"Rated false" × warning	0.03
	(0.05)
Constant	1.85***
	(0.03)
Question fixed effects	Yes
N (responses)	11962
Respondents	2554

OLS models with robust standard errors clustered by respondent. "Accuracy" measures belief that a headline is accurate from "Not at all accurate" (1) to "Very accurate" (4). Respondents in the pure control condition are excluded, as are responses from participants in the disputed and false conditions who saw a false headline that did not include a "Disputed" or "Rated false" tag

*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$ (two-sided)

We find no support for H3. Our results do not allow us to reject the null hypothesis of no difference in the effect of the "Disputed" and "Rated false" tags when a general warning is present compared to when it was not. The marginal effects of the tags remain negative and statistically significant when a general warning was previously provided, however (disputed: -0.20, $p < 0.01$; false: -0.31, $p < 0.01$). We therefore conclude that a general warning did not augment the effect of tags warning about specific misleading articles. Participants view specific warning tags immediately before answering questions about a headline's accuracy, so the tag's effect may be more immediate and take precedence over that of the general warning shown earlier.

Differences by Article Slant

We next test whether these effects vary depending on whether the general or specific warnings provided to participants are politically congenial or uncongenial. To do so, we add a directional preference measure (an indicator for Trump approval) and corresponding interaction terms to our previous statistical model predicting the perceived accuracy of false news headlines. The coefficients of interest are presented in Table 4, which separately estimates results for false pro- and anti-Trump articles.

Table 4 Experimental effects on perceived accuracy of false headlines by article slant

	Anti-Trump	Pro-Trump
General warning	− 0.11**	− 0.06
	(0.05)	(0.05)
"Disputed" tag	− 0.28***	− 0.24***
	(0.05)	(0.05)
"Rated false" tag	− 0.41***	− 0.37***
	(0.05)	(0.05)
Trump approval	− 0.18***	0.64***
	(0.07)	(0.07)
Warning × Trump approval	− 0.04	0.06
	(0.09)	(0.10)
"Disputed" × warning	0.10	0.04
	(0.07)	(0.07)
"Disputed" × Trump approval	0.14	0.03
	(0.09)	(0.10)
"Disputed" × warning × Trump approval	− 0.05	− 0.12
	(0.13)	(0.15)
"Rated false" × warning	0.12*	0.01
	(0.07)	(0.07)
"Rated false" × Trump approval	0.23**	− 0.01
	(0.10)	(0.10)
"Rated false" × warning × Trump approval	− 0.06	− 0.07
	(0.13)	(0.15)
Constant	1.91***	1.99***
	(0.04)	(0.04)
Question fixed effects	Yes	Yes
N (responses)	5972	5972
Respondents	2550	2548

OLS models with robust standard errors clustered by respondent. "Accuracy" measures belief that a headline is accurate from "Not at all accurate" (1) to "Very accurate" (4). Respondents in the pure control condition are excluded, as are responses from participants in the disputed and false conditions who saw a false headline that did not include a "Disputed" or "Rated false" tag

*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$ (two-sided)

These results are estimated only among respondents who approve or disapprove of Trump.[17]

Though Trump job approval is strongly associated with baseline levels of belief in the false headlines, our findings do not support our hypothesis that people would

[17] A typo in the pre-registration statement to this effect instead mistakenly stated we would exclude "pure independents." The results below again exclude pure controls but equivalent results including those respondents are provided in Online Appendix C. We do not include respondents with no opinion of Trump in that model because there were so few ($n = 4$).

resist warnings about politically congenial false news (H4). We also find no difference in the effect of a general warning, a research question for which we had weaker expectations because it does not challenge a specific story like a fact-checking tag (RQ1). For pro-Trump stories, we are unable to reject the null hypothesis that the effects of the "Disputed" or "Rated false" tags do not differ between Trump approvers and disapprovers. We also find no measurable difference in the effect of the "Disputed" tags by Trump approval for anti-Trump stories. One possible explanation for these findings is that the headlines we tested were likely unfamiliar to many respondents. As a result, they may have been less connected to respondents' personal beliefs and thus more easily debunked than well-known misperceptions.[18]

In one case, the effect of a "Rated false" tag did significantly differ between Trump approvers and disapprovers for anti-Trump stories, but the sign of the effect was the opposite of the hypothesized direction. The presence of a "Rated false" tag reduced the perceived accuracy of anti-Trump headlines significantly *more* among Trump disapprovers for whom the headlines were politically congenial than among Trump approvers for whom they were uncongenial (0.20, $p < 0.05$). This counterintuitive finding may be the result of differing levels of trust in fact-checking and false news susceptibility. An exploratory analysis using pre-treatment measures shows that Trump approvers were much more likely to believe in election 2016 false news (an average of 0.26 points higher on a four-point accuracy scale, $p < 0.01$) and to distrust fact-checkers (an average of 0.53 points lower on a four-point trust scale, $p < 0.01$) than disapprovers.

These results can be observed in Fig. 2, which displays the mean level of perceived accuracy by condition for both anti- and pro-Trump headlines among respondents who did not receive a general warning. In general, the reduction in belief is similar regardless of whether the headline is politically congenial. The exception was anti-Trump headlines, which Trump disapprovers rate as more accurate than approvers when no tag is present, but view as less accurate when a "Rated false" tag accompanies the headline.[19]

Spillover Effects

Finally, we consider our last two research questions, which concern possible unintended spillover effects from general or specific warnings. We first test whether the

[18] Pennycook and Rand (2018a) similarly find that "the correlation between CRT [Cognitive Reflection Test scores] and perceived accuracy is unrelated to how closely the headline aligns with the participant's ideology... Our findings therefore suggest that susceptibility to fake news is driven more by lazy thinking than it is by partisan bias per se." Similarly, Porter et al. (2018) find minimal differences between ideological groups in their willingness to accept false news headlines.

[19] We conducted an additional exploratory analysis to test whether the effects of political congeniality were altered by a participant's political knowledge. Consistent with previous research, we found that high political knowledge was associated with a lower belief in false news stories regardless of the article's slant. However, we did not find convincing evidence that high political knowledge meaningfully changed a specific warning's effect on belief in false news headlines. Results for this exploratory analysis are included in Online Appendix C (Table C16).

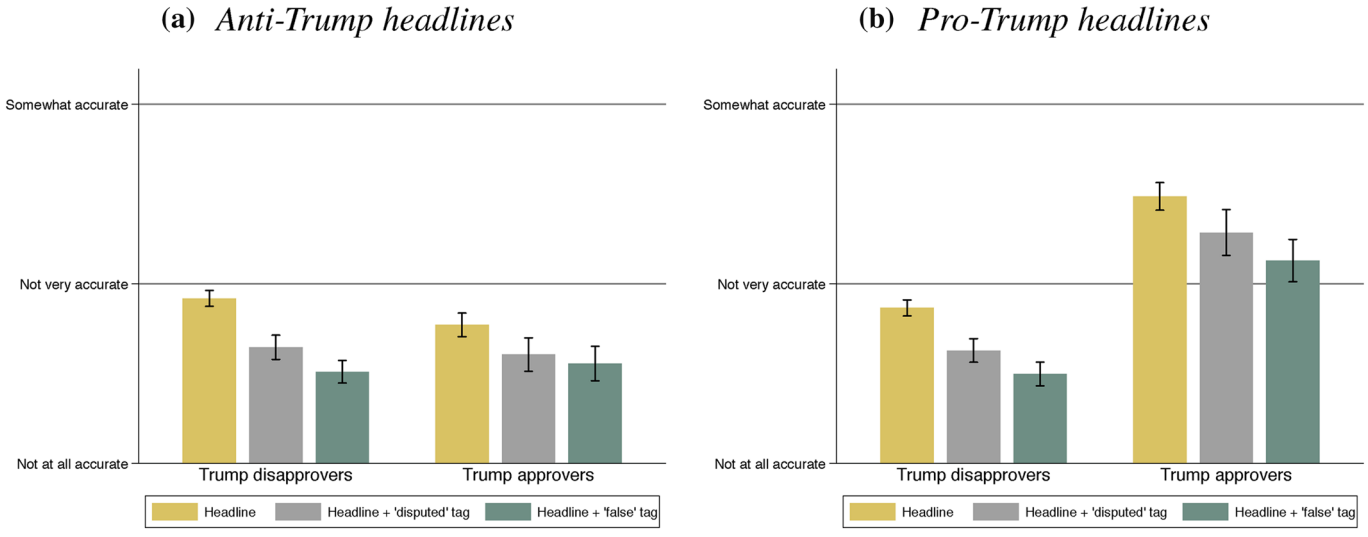

Fig. 2 Specific warning effects by political congeniality of false headlines. Mean belief that false headlines were accurate on a four-point Likert scale from "Not at all accurate" (1) to "Very accurate" (4). See Online Appendix A for question wording and stimulus materials. Excludes respondents assigned to receive a general warning as well as those in the pure control condition

Table 5 Experimental tests for spillover effects of warnings on perceived accuracy

	Untagged false headlines	True news headlines
General warning	− 0.08**	− 0.12***
	(0.03)	(0.04)
"Disputed" condition	0.02	0.06
	(0.05)	(0.04)
"Rated false" condition	0.04	0.03
	(0.04)	(0.04)
"Disputed" × warning	0.06	0.09
	(0.06)	(0.06)
"Rated false" × warning	0.03	0.14**
	(0.06)	(0.06)
Constant	1.87***	2.87***
	(0.03)	(0.03)
Question fixed effects	Yes	Yes
N (responses)	7968	6585
Respondents	2436	2502

OLS models with robust standard errors clustered by respondent. "Accuracy" measures belief that a headline is accurate from "Not at all accurate" (1) to "Very accurate" (4). Respondents in the pure control condition are excluded, as are responses from participants in the disputed and false conditions who saw a "Disputed" or "Rated false" tag for the specific headline in question

*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$ (two-sided)

presence of "Disputed" or "Rated false" tags affects belief in untagged false or true news headlines (RQ3). In this case, the presence of tags could cause a contrast effect that leads participants to infer that other stories are accurate. Alternatively, the presence of these tags could make participants more skeptical about all news. We therefore test whether a general warning about false news affects the perceived accuracy of true news headlines (RQ3).

These research questions are evaluated in Table 5, which separately tests these effects on the perceived accuracy of both types of headlines. The models reported in the first column exclude headlines in which respondents in the "Disputed" or "Rated false" conditions saw a tag; they thus test the effect of assignment to those conditions on the perceived accuracy of *other* headlines.[20]

We find no statistically measurable effect on the perceived accuracy of untagged false articles when other false articles had a "Disputed" or "Rated false" tag (RQ2a). However, our point estimates for "Disputed" (0.02) and "Rated false" (0.04) are similar to the "implied truth" effect found by Pennycook and Rand (2017) (0.03); we simply lack the precision to detect an effect of that magnitude (95% CI for

[20] Headlines viewed by respondents in the "Disputed" or "Rated false" conditions before exposure to the first tag are also excluded (spillover is impossible for participants who are not yet treated).

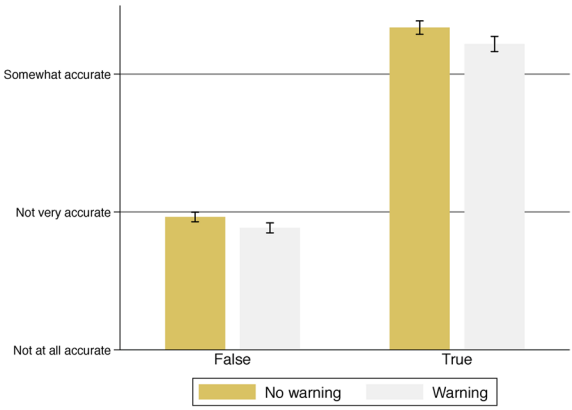

Fig. 3 General warning effects on belief in true and false news articles. Mean belief that true and false news headlines were accurate on a four-point Likert scale from "Not at all accurate" (1) to "Very accurate" (4). See Online Appendix A for question wording and stimulus materials

"Disputed": −0.07, 0.11; 95% CI for "Rated false": −0.05, 0.16). The perceived accuracy of true articles was also unaffected when other false articles included these tags (RQ2b).

However, the general warning had an unintended spillover effect on the perceived accuracy of true headlines (− 0.12, $p < 0.01$), though the substantive magnitude of this effect was small (Cohen's d = 0.12). While overall levels of belief in false news stories were much lower than for true ones (Fig. 3), the decrease in perceived accuracy from the general warning was greater with true stories (− 0.12) than false ones (− 0.08), though an exploratory analysis pooling evaluations of true and false headlines shows we cannot reject the null hypothesis of no difference in perceived accuracy between them. The findings suggest that the specific warnings were more effective because they reduced belief solely for false headlines and did not create spillover effects on perceived accuracy of true news.

Conclusion

This study provides several important new findings about how to most effectively counter false information on social media. First, both "Disputed" and "Rated false" tags modestly reduce belief in false news. Notably, we find larger accuracy effects for the "Disputed" tags than Pennycook and Rand 2017.[21] However, our results demonstrate that "Rated false" tags, which specifically tell users when claims made in headlines are untrue, are more effective at reducing belief in misinformation than the "Disputed" tags previously used by Facebook. Encouragingly, we find no consistent evidence that the effects of these tags varies by the political congeniality of the headlines or that exposure to the tags increases the perceived accuracy of

[21] This difference in effect size could be partially attributable to respondents being aware that their ability to discern true from false headlines was under scrutiny, since they had previously been asked to rate political statements as true or false at the beginning of our survey.

unlabeled false headlines (though our study lacks the precision necessary to detect the small "implied truth" effect that Pennycook and Rand 2017 identify).

By contrast, though general warnings about false news also appear to decrease belief in false headlines, the effect of a general warning is small compared to either type of tag. Moreover, general warnings also reduce belief in real news and do not enhance the effects of the "Rated false" and "Disputed" tags, suggesting that they are a less effective approach.

Our results provide support for prior studies finding a negative effect of general warnings on belief in misinformation (Bolsen and Druckman 2015; Ecker et al. 2010; van der Linden et al. 2017), but our finding that these warnings also reduce the perceived accuracy of true headlines suggest that they pose a potential hazard. False news may already increase distrust in legitimate information; unintended spillover effects from general warnings or related proposals to fight false information by increasing media literacy (e.g., Atkins 2017) could exacerbate this problem. Our "Disputed" and "Rated false" tags, which more effectively reduce the perceived accuracy of false headlines without causing these spillover effects, may be a safer way to reduce belief in misinformation.

Further research is needed to evaluate this finding and better understand the mechanism for the spillover effect we observe. One potential explanation is that warnings about false news prime people to think about misleading information online, making them less likely to trust any articles they see on social media. Another possible interpretation is that we are observing a "tainted truth" effect in the context of political misinformation. In social cognition research, such an effect occurs when eyewitnesses who are warned about the influence of misinformation overcorrect for this threat and identify fewer true items than eyewitnesses who are not warned (Echterhoff et al. 2007; Szpitalak and Polczyk 2010). Our results suggest that the tainted truth effect could apply to prospective warnings about political misinformation, but further research is necessary to test if this finding holds in other contexts and designs.

Our study has several important limitations that should be addressed in future research. First, our sample was more educated and politically active than the general population and leaned liberal; future studies should be conducted with nationally representative sample. A second limitation is that our study examined the effect of warnings and tags on pro- and anti-Trump headlines in the aftermath of a presidential election—a substantively important but specific context. While Trump-related headlines remain timely and salient, further research should seek to determine whether our results hold in other contexts and for other types of false headlines. Third, we do not examine over-time effects. Future studies should evaluate long-term belief in false news after the initial exposure and how our manipulations affect those beliefs. Fourth, as noted above, our design does not allow us to identify the causal mechanisms responsible for the effects we observed—a challenge facing nearly all experimental studies (Bullock et al. 2010). In particular, future research should employ designs that provide more leverage for understanding the effects of warnings on belief in false news. Finally, as in any experimental study, we cannot fully rule out the possibility of demand effects. Any survey that asks about the perceived accuracy of political statements and the effects of interventions on those

self-reports is susceptible to these effects, though research suggests they are rare (Mummolo and Peterson 2018).

Our study also made a number of design choices that should be revisited in future research. First, we focused on belief in headlines because they are prominently displayed on social media, but future studies should also measure the effects of warnings and tags on belief when people actually read the articles in question. Second, false news articles typically appear on a user's timeline because a friend liked or shared the article, but we chose not to test the effect of social endorsements or other contextual cues on belief in false news articles. Future research should use field experiments or conduct studies in other settings to evaluate the extent to which our findings generalize to real-world contexts. Third, our headlines omitted article sources to allow us to isolate the effect of our treatments, but these sources are likely to play a role in how individuals evaluate Facebook posts. Exploring the interactions between source credibility and warnings on belief in misinformation is another important avenue for future research (though see Pennycook and Rand 2018b and Clayton et al. 2018a, who both find that source cues may play a limited role in credibility assessments of true and false news). Finally, we chose to use the same two fact-checking sources throughout the study, PolitiFact and Snopes. However, people differ in how much they trust the most prominent national fact-checking organizations (e.g., Nyhan and Reifler N.d.). Future studies should vary the source of fact-checks in order to determine whether the fact-checking source influences individuals' perceptions of true and false headlines.

Despite these limitations, this study provides important insights into how efforts to prevent belief in misinformation on social media could be more effective and suggests that online false news can be countered with some degree of success.

Acknowledgements We thank the Dartmouth College Office of Undergraduate Research for generous funding support. We are also grateful to Ro'ee Levy and David Rand for helpful comments.

References

Aklin, M., & Urpelainen, J. (2014). Perceptions of scientific dissent undermine public support for environmental policy. *Environmental Science and Policy*, *38*, 173–177.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Atkins, L. (2017). States should require schools to teach media literacy to combat fake news. The Huffington Post, July 13, 2017. Retrieved July 8, 2018, from https://www.huffingtonpost.com/entry/states-should-require-schools-to-teach-media-literacy_us_59676573e4b07b5e1d96ed86.

Berinsky, A. J., Gregory, A. H., & Gabriel, S. L. (2012). Evaluating online labor markets for experimental research: Amazoncom's mechanical turk. *Political Analysis*, *20*(3), 351–368.

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, *65*(4), 619–638.

Bolsen, T., & Druckman, J. N. (2015). Counteracting the politicization of science. *Journal of Communication*, *65*(5), 745–769.

Boykoff, M. T., & Boykoff, J. M. (2004). Balance as bias: Global warming and the US prestige press. *Global Environmetal Change*, *14*, 125–136.

Bullock, J. G., Green, D. P., & Shang, E. H. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*(4), 550–558.

Chinn, S., Lane, D. S., & Hart, P. S. (2018). In consensus we trust? Persuasive effects of scientific consensus communication. *Public Understanding of Science*, *27*(7), 807–823.

Clayton, K., Davis, J., Hinckley, K., & Horiuchi, Y. (2018). *Partisan motivated reasoning and misinformation in the media: Is news from ideologically uncongenial sources more suspicious*? Unpublished manuscript. Retrieved July 8, 2018, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035272.

Constine, J. (2017). Facebook puts link to 10 tips for spotting 'false news' atop feed. Tech Crunch, April 6, 2017. Retrieved July 18, from https://techcrunch.com/2017/04/06/facebook-puts-link-to-10-tips-for-spotting-false-news-atop-feed/.

Coppock, A. (2016). *Generalizing from survey experiments conducted on mechanical Turk: A replication approach*. Unpublished manuscript, March 22, 2016. Retrieved July 8, 2018, from https://alexandercoppock.files.wordpress.com/2016/02/coppock_generalizability2.pdf.

Corbett, J. B., & Durfee, J. L. (2004). Testing public (un)certainty of science media representations of global warming. *Science Communication*, *26*(2), 129–151.

de Leeuw, E. D., Hox, J. J., & Boevé, A. (2015). Handling do-not-know answers: Exploring new approaches in online and mixed-mode surveys. *Social Science Computer Review*, *34*(1), 116–132.

Echterhoff, G., Groll, S., & Hirst, W. (2007). Tainted truth: Overcorrection for misinformation influence on eyewitness memory. *Social Cognition*, *25*(3), 367–409.

Ecker, U. K. H., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied*, *20*(4), 323–335.

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition*, *38*(8), 1087–1100.

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, *38*(S1), 127–150.

Gabielkov, M., Ramachandran, A., Chaintreau, A., Legout, A. (2016). Social clicks: What and who gets read on twitter? In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*. Ne York: ACM.

Gottfried, J., & Shearer, E. (2017). *News use across social media platforms 2016*. Pew Research Center, May 26, 2016. Retrieved May 23, 2017, from http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425.

Kahan, D. M. (2015). Climate-science communication and the measurement problem. *Political Psychology*, *36*(S1), 1–43.

Kahan, D. M., Dawson, E. C., Peters, E., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, *1*(1), 54–86.

Koehler, D. J. (2016). Can journalistic false balance distort public perception of consensus in expert opinion? *Journal of Experimental Psychology: Applied*, *22*(1), 24–38.

Krupnikov, Y., & Levine, A. S. (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science*, *1*(1), 59–80.

Kuru, O., Pasek, J., & Traugott, M. W. (2017). Motivated reasoning in the perceived credibility of public opinion polls. *Public Opinion Quarterly*, *81*(2), 422–446.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Manjoo, F. (2013). You won't finish this article. *Slate Magazine*, June 6, 2013. Retrieved May 23, 2017, from http://www.slate.com/articles/technology/technology/2013/06/how_people_read_online_why_you_won_t_finish_this_article.html.

Mosseri, A. (2016). *Addressing Hoaxes and Fake news*. Facebook, December 15, 2016. Retrieved July 18, 2018, from https://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

Mosseri, A. (2017). *A new educational tool against misinformation*. Facebook, April 6, 2017. Retrieved May 23, 2017, from https://newsroom.fb.com/news/2017/04/a-new-educational-tool-against-misinformation/.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138.

Mummolo, J., & Peterson, E. (2018). Demand effects in survey experiments: An empirical assessment. *American Political Science Review.* Retrieved January 5, 2019, from https://scholar.princeton.edu/sites/default/files/jmummolo/files/demand_effects_in_survey_experiments_an_empirical_assessment.pdf.

Nyhan, B., Porter, E., Reifler, J. & Wood, T. (2017). *Taking corrections literally but not seriously? The effects of information on factual beliefs and candidate favorability*. Unpublished manuscript. Retrieved July 8, 2018, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2995128.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Nyhan, B., & Reifler, J. (N.d). *Do people actually learn from fact-checking? Evidence from a longitudinal study during the 2014 campaign*. Unpublished manuscript. Retrieved September 20, 2017, from http://www.dartmouth.edu/~nyhan/fact-checking-effects.pdf.

Oremus, W. (2017). Facebook has stopped saying fake news. Slate, August 8, 2017. Retrieved July 8, 2018, from http://www.slate.com/blogs/future_tense/2017/08/08/facebook_has_stopped_saying_fake_news_is_false_news_any_better.html.

Owen, L. H. (2018). *Is your fake news about immigrants or politicians? It all depends on where you live*. Nieman Journalism Lab, May 25, 2018. Retrieved July 18, 2018, from http://www.niemanlab.org/2018/05/is-your-fake-news-about-immigrants-or-politicians-it-all-depends-on-where-you-live/.

Pennycook, G., & Rand, D. G. (2017). *The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings*. Unpublished manuscript. Retrieved July 8, 2018, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384.

Pennycook, G., & Rand D. G. (2018a). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition.* Retrieved July 8, 2018, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3165567.

Pennycook, G., & Rand, D. G. (2018b). *Who falls for fake news? The roles of analytic thinking, motivated reasoning, political ideology, and bullshit receptivity*. Unpublished manuscript. Retrieved July 10, 2018, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3023545.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2017). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*. Retrieved May 23, 2017, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2958246.

Porter, E., Wood, T. J., & Kirby, D. (2018). Sex trafficking, Russian infiltration, birth certificates, and pedophilia: A survey experiment correcting fake news. *Journal of Experimental Political Science*, *5*(2), 159–164.

Schaedel, S. (2017). Black lives matter blocked hurricane relief? Factcheck.org, September 1, 2017. Retrieved September 26, 2017 from http://www.factcheck.org/2017/09/black-lives-matter-blocked-hurricane-relief/.

Silverman, C. (2016). *This analysis shows how viral fake election news stories outperformed real news on facebook*. Buzzfeed, November 16, 2016. Retrieved May 22, 2017, from https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook?utm_term=.lgQvmj974#.qqXqL1AJV.

Silverman, C., & Jeremy S.-V. (2016). *Most Americans who see fake news believe it, new survey says*. December 6, 2016. Retrieved May 23, 2017, from https://www.buzzfeed.com/craigsilverman/fake-news-survey?utm_term=.srvopPEVR#.cqlAz0PeX.

Smith, J., Jackson, G., & Raj, S. (2017). *Designing against misinformation*. Medium, December 20, 2017. Retrieved July 8, 2018, from https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2.

Szpitalak, M., & Polczyk, R. (2010). Warning against warnings: Alerted subjects may perform worse. Misinformation, involvement and warning as determinants of witness testimony. *Polish Psychological Bulletin*, *41*(3), 105–112.

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*(3), 460–480.

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008.

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe Report, September 27, 2017. Retrieved July

8, 2018, from https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Wegner, D. M., Richard Wenzlaff, R., Kerker, M., & Beattie, A. E. (1981). Incrimination through innuendo: Can Media questions become public answers? *Journal of Personality and Social Psychology*, *40*(5), 822–832.

Affiliations

Katherine Clayton[1] · Spencer Blair[1] · Jonathan A. Busam[1] · Samuel Forstner[1] · John Glance[1] · Guy Green[1] · Anna Kawata[1] · Akhila Kovvuri[1] · Jonathan Martin[1] · Evan Morgan[1] · Morgan Sandhu[1] · Rachel Sang[1] · Rachel Scholz-Bright[1] · Austin T. Welch[1] · Andrew G. Wolff[1] · Amanda Zhou[1] · Brendan Nyhan[2]

Katherine Clayton
Katherine.P.Clayton.GR@dartmouth.edu

[1] Dartmouth College, Hanover, NH, USA

[2] University of Michigan, Ann Arbor, MI, USA