Correcting What's True:

Testing Competing Claims about Health Misinformation on Social Media

By:

Emily K. Vraga

University of Minnesota

Leticia Bode

Georgetown University

Accepted for publication in the *American Behavioral Scientist.*

Vraga is an Associate Professor and is the Don and Carole Larson Professor in Health Communication in the Hubbard School of Journalism and Mass Communication at the University of Minnesota. Correspondence about this manuscript should be addressed to her at 338 Murphy Hall, 206 Church Street, Minneapolis, MN 55455.

Bode is a Provost's Distinguished Associate Professor in the Communication, Culture, and Technology master's program at Georgetown University at 3520 Prospect st NW Suite 311, Washington DC 20057.

Correcting What's True:

Testing Competing Claims about Health Misinformation on Social Media

Abstract

This study expands on existing research about correcting misinformation on social media. Using an experimental design, we explore the effects of three truth signals related to stories shared on social media: whether the person posting the story says it is true, whether the replies to the story say it is true, or whether the story itself is actually true. Our results suggest that individuals should not share misinformation in order to debunk it, as audiences assume sharing is an endorsement. Additionally, while two responses debunking the post do reduce belief in the post's veracity and argument, this process occurs equally when the story is *false* (thereby reducing misperceptions) as when it is *true* (thus creating misperceptions). Our results have implications for individuals interested in correcting health misinformation on social media and for the organizations who support their efforts.

Correcting What's True:

Testing Competing Claims about Health Misinformation on Social Media

Social media is awash with health misinformation (Avaaz, 2020; Mitchell, Gottfried, Stocking, Walker, & Fedeli, 2019; Wilner & Holton, 2020), which raises particular dangers because it can lead to health behaviors that increase personal and public risk (Bigsby & Hovick, 2018; Jolley & Douglas, 2014).

We contribute to research on health misinformation on social media in three major ways. First, most existing work on correcting misinformation on social media has focused on the role that people or groups play in *responding* to misinformation being shared by another person or group (e.g., Bode & Vraga, 2018; Margolin et al., 2018; Vraga & Bode, 2017). But roughly 5% of the American public (Mitchell et al., 2019) reports that they have intentionally shared misinformation to call attention to its falseness. Therefore, in this study we test whether sharing misinformation while explicitly saying it is false effectively debunks it.

Second, we contrast the role of *posting* (even to debunk) misinformation with the role of *responding* to misinformation. In general, when people see someone else sharing misinformation experience correction from algorithms, experts, or other people on social media, they view the post as less credible and reduce their misperceptions on the issue – a process termed *observational correction* (Bode & Vraga, 2015; Vraga & Bode, 2017). However, this research has only considered people responding to someone *endorsing* misinformation, rather than what occurs when an original poster is attempting to *debunk* the misinformation. In this study, we compare effects when people respond to debunk versus validate the information's veracity, and how this relates to the position of the original person sharing the story (whether they say the story is true or false) to explore the relative power of the original poster and the responses in

affecting beliefs. While belief in the position promoted by the story should be lowest when both the original poster and those responding to the post agree a story is false – and highest when all people agree it is true – it is less clear what happens when the original poster and those responding disagree about the veracity of the information, allowing us to test two-sided message flows (Zaller, 1992) in the context of misinformation and correction.

Third, previous work on observational correction has focused on correcting stories that are, in fact, misinformation. However, misinformation is a complicated construct, which can include the type of content, the motive of the creator, and how the information is shared (e.g., Molina, Sundar, Le, & Lee, 2021; Tandoc, Lim, & Ling, 2018; Wardle, 2017). Importantly, however, we also need to consider whether the mechanisms of observational correction work for *true* stories as well. That is, can someone's beliefs about an issue become *less* well-informed, based on seeing someone "correct" a *true* story with *false* information? To test this, we include two equally implausible stories, one of which is true and one of which is false.

This study thus tests several different signals that people may use to determine whether something they see on social media is true or not. Specifically, we manipulate three truth signals: 1) whether the poster says the story is true or false, 2) whether the replies say the story is true or not, and 3) whether the story is *actually* true or false. We consider the effects of these three truth signals on three related but distinct outcomes: perceptions of story veracity, attitude alignment with the position advocated by the story, and related behavioral intentions. Each of these outcomes has been studied in previous research (e.g., Bode & Vraga, 2015; Smith & Seitz, 2019; van der Meer & Jin, 2020; Vraga et al., 2020; Walter et al., 2020) and collectively speak to the potential impact of exposure to (mis)information and correction on social media, allowing us to

better understand how correction functions in the wild world of social media, where competing truth signals may conflict.

Literature review

Sharing to Debunk

The first element we consider is the information provided by the poster with regards to the veracity of the information they are sharing. Most research on misinformation and correction on social media assume that those sharing the misinformation implicitly or explicitly endorse its inaccurate claims (e.g., Bode & Vraga, 2018; Smith & Seitz, 2019; Vraga et al., 2020). This does not have to be the case. In one study, of the 10% of U.S. adults who admitted to sharing misinformation they knew at the time was made up, a plurality (46%) said they did so to tell other people that the information was inaccurate (Mitchell et al., 2019).

This raises the question of whether sharing inaccurate stories to debunk misinformation is effective. On the one hand, the original poster can frame the issue in their own terms by suggesting whether the story is true or false (Chong & Druckman, 2007). Indeed, one theoretical reason for the continued influence effect is that misinformation occurs first, creating a primacy effect that endures in the face of later corrections (Ecker et al., 2011). Moreover, the implicit endorsement of someone sharing a news story on social media can affect trust (Turcotte et al., 2015), and these poster cues are especially important when the story comes from a distrusted news source (Oeldorf-Hirsch & deVoss, 2020), as it does here. In general, reports that a story is untrue tend to reduce story veracity and misperceptions (e.g., Porter & Wood, 2019; Walter & Murphy, 2018), which we expect to also occur when an original poster says a story is false.

On the other hand, people may not pay much attention to what the poster says. At least one study suggests that Twitter users largely retweet content that agrees with their position – and

points out that "retweets are not endorsements" statements signal concerns that sharing content *is* seen as an endorsement (Metaxas, et al., 2015; see also Molyneux & Mourao, 2019). Additionally, people's attention is often drawn to visual elements and pictures on social media (Lee & Shin, 2021; Vraga et al., 2016), so people may overlook the original poster's comments to focus on the story itself and its headline.

This latter research raises the question to what degree people recall the position of someone posting a story on social media. We begin by asking whether people recall the original poster's position, and whether recall depends on whether the poster says the story is true or false, allowing us to test the assumption that a posting signals endorsement.

RQ1: To what extent do people recall the original poster's position with regards to the story, and does it differ depending on whether they say the story is true or false?

Although we do not know the amount of attention paid to the original poster cue, which could undermine its effect on attitudes, broadly speaking we believe the literature supports the argument that the original poster's position about the story should predict audience perceptions.

H1: When the original poster says a story is false, (a) assessments of story veracity, (b) issue attitudes, and (c) behavioral intentions will be lower than when replies say the story is true.

Responding to Debunk

Second, we consider the success of corrective *responses* to misinformation shared on social media. Although many people ignore misinformation on social media, others actively correct it (Tandoc, Lim, & Ling, 2020). A robust literature demonstrates that responsive correction from experts (van der Meer & Jin, 2020; Vraga & Bode, 2017), other people on social media (Bode & Vraga, 2018; Margolin et al., 2018; Vraga et al., 2020), and platforms

themselves (Bode & Vraga, 2015; Smith & Seitz, 2019) reduce information credibility and misperceptions. Indeed, a meta-analysis found corrections in response to health misinformation on social media significantly reduced misperceptions (Walter et al., 2020). When these corrections come from other individuals, best practices suggest that multiple people, citing credible sources using a link, are likely to be most effective (Vraga & Bode, 2020). As our responsive corrections follow these best practices, we expect:

H2: When replies say a story is false, (a) assessments of story veracity, (b) issue attitudes, and (c) behavioral intentions will be lower than when replies say the story is true.

To our knowledge no research has tested *recall* for these replies. Therefore, we also consider the degree to which people recall the position of the reply posts, and whether recall differs depending on whether the replies say the message is true or false.

RQ2: To what extent do people recall the replies' position with regards to the story, and does it differ depending on whether they say the story is true or false?

Considering the Poster and the Responses Together

While previous work regarding observational correction inherently deals with disagreement, as they test what occurs when others respond to misinformation shared (and explicitly or implicitly endorsed) by another individual (Bode & Vraga, 2018; Smith & Seitz, 2019; Vraga & Bode, 2017), these studies do not test how these forces – the position of the original poster and of the responses with regards to the veracity of the story – interact. Message effects are strongest when all sources are in agreement in position or framing, and weaker when competition exists (Chong & Druckman, 2007; Druckman, 2004). This scenario represents a

"one-sided" information flow, so we expect individuals to come into alignment with the common message being shared (Zaller, 1992).

H3: An interaction between original poster and replies exists, such that (a) assessments of story veracity, (b) issue attitudes, and (c) behavioral intentions are lowest when both say the story is false, and highest when both say the story is true.

Message effects are weaker when people enter a "two-sided" information flow or "competitive" framing environment, especially when pre-existing attitudes on the issue are likely to be relatively weak (Chong & Druckman, 2007; Druckman, 2004; Zaller, 1992). However, we are unsure whether the position of the *original poster* or those *responding* will be given more weight when such disagreement occurs. The original poster provides the first "frame" and may set initial expectations about the veracity of the information. However, this does not mean that people actually *view* the original poster's content first. Indeed, one eye tracking study found that attention to a post is highest when the original poster and the comments take opposing positions (Suflow, Schafer, & Winter, 2018), but does not distinguish attention to the original post versus comments. In this case there are multiple comments, relying on expert sources to debunk the original post (Vraga & Bode, 2020; Walter et al., 2020). Therefore, they may be given more weight than the original poster, who offers little evidence to substantiate their claim that the story is true or false. Because the literature does not suggest clear expectations, we ask:

RQ3: How are (a) assessments of veracity, (b) issue attitudes, and (c) behavioral intentions affected when the original poster and the replies differ in their assessment of story veracity?

The Role of Truth

Finally, and of critical importance, we consider how these processes differ when the story being shared is *actually true* versus when it is in fact misinformation. Several signals can help people identify whether a story is true or false – notably, the position of the original poster and the responses to the story, as we outline above. Additionally, there are signals *within* the story itself: for example, the "proximate" source of the story (i.e., the organization that posted the story originally; Kang et al., 2011) and perceptions of its general plausibility (i.e., whether or not people believe the story is likely to be true; Lee & Shin, 2021). As we detail below, we hold these latter two signals constant, by having the stories come from the same low credibility source (NationalReport.net) and choosing stories considered equally implausible in a pre-test. This allows us to test whether the effects of the original poster position and reply position depend on the *actual accuracy* of the story: one story is true and the other is false.

Existing research has confirmed that observational correction works when a false story is corrected (Bode & Vraga, 2018; Smith & Seitz, 2019; Vraga & Bode, 2017), and we theoretically argue this process should also occur when the original poster says it is false. Because we have deliberately kept the signals regarding the veracity of the story constant, it is possible that people will respond equally to the truth signals from the original poster and the replies regardless of whether the story is true or false. This would pose serious questions for how and when we encourage people to engage in such corrections. We therefore ask:

> *RQ4: Do the effects of original poster and reply position on (a) assessments of story credibility, (b) issue attitudes, and (c) behavioral intentions differ when the original story is true versus when it is false?*

Pre-Test

To identify common health misinformation on social media, we reviewed the "Fact Check Archive" for Medical Fact-checks on Snopes.com in Summer 2017. We selected medical stories that had been fact-checked or updated in the previous year (between 2016 and 2017). From that, we selected two stories rated as "true" and 6 stories rated "false" for pre-testing.[1] We use Snopes.com because it has been the basis for identifying misinformation in previous work (e.g., Allcott & Gentzkow, 2017; Margolin et al., 2018) and is the most prominent and least polarized fact-checking website among young adults (Knight/Gallup, 2018).

For the pretest, 100 participants from Amazon's Mechanical Turk[2] were shown headlines from each of the eight Snopes stories in random order. For each story, they were asked whether they had seen or heard anything about the story previously, their rating of story veracity,[3] their certainty in that evaluation, and their rating of the headline on five-point bipolar scales from *implausible-plausible, believable-not believable, likely to be true-likely to be false*. We report the descriptive statistics for the selected headlines in Table 1. Based on this pre-test, we selected one true ("Scientists discover a 'zombie tick' whose bite makes you allergic to red meat," which we refer to as ticks) and one false ("Scientists discover that public water fluoridation is reducing public intelligence," which we refer to as fluoride)[4] headline. Notably, both stories were seen as relatively implausible.

Methods

[1] Only a small minority of the Medical Fact-Checks on Snopes were rated as true, necessitating this imbalance in true versus false stories.
[2] Participants were 59% male, educated (67% of the sample had at least a Bachelor's degree), and Democratic (56% Democrats, 22% Independent, 22% Republican).
[3] This item was measured on a five-point scale, from definitely true to definitely false. An additional response option was "unsure," which was included in the mixed (both true and false) middle category for analyses.
[4] While scientists agree that fluoride does *not* lower IQ, very recent information (published after our data collection) suggests there may be an effect of fluoride consumed by pregnant women on the IQ of their fetuses (Harris, 2019).

Data for the study were collected during the Fall 2017 and Spring 2018 semesters at a large mid-Atlantic university. Participants from the university-required basic communication course at a large, diverse public university received course credit to participate. The data were reviewed to remove duplicate responses, participants who did not consent, or who reported they were under the age of 18 or over the age of 60, leaving a total sample of 1431. Participants were an average of 19 years old (M=19.45, $S.D.$=2.87, min=18, max=54), slightly more female (55% female, 45% male), racially diverse (46% white, 22% Asian-American, 14% African-American, and 11% Hispanic), and largely in their first year in college (62% freshman).

In this study, we analyze a 2 (Poster position: True vs false) x 2 (Reply position: True vs False) x 2 (Article accuracy: True vs False) experimental design, extracted from a larger design (N for these eight cells=821).[5]

After a short questionnaire, participants were exposed to a simulated Facebook News Feed, which they were told was taken from someone's feed, and asked to read it as if it were their own feed. Participants viewed five posts, with the experimental manipulation embedded in the second post on the page (see Appendix[6]). Our first manipulation was article accuracy – the ticks (true) story versus the fluoride (false) story. In all conditions, the manipulated story appeared to come from NationalReport.Net, which Wikipedia lists as a fake news website (Wikipedia.org, 2019).

The second manipulation involved the position of the person *sharing* the story. In the "poster true" condition, the poster reinforced the story headline, either claiming "Ticks can cause meat allergies" or "fluoride in drinking water reduces IQ." In the "poster false" condition, the

[5] Additional cells not analyzed in this study are not crossed with the experimental factors examined here so are excluded from all analyses (N=610).
[6] Supplemental appendix available at: https://doi.org/10.7910/DVN/9Y0SSL

poster labels the story as "bogus" and explicitly debunks the claim (e.g. "check out this bogus story: Ticks can't cause meat allergies!").

The third manipulation involves the position of the people *responding* to the story. Using best practices, two individuals responded to the original post, providing links to fact-checking stories on the topic (Vraga & Bode, 2020) from Snopes.com and USAToday.com. The fact check headlines are inconclusive in both conditions. In the "replies true" conditions, the replies claim the story is "completely true" and "confirmed" by the fact checks. In the "replies false" conditions, the replies claim the story is "completely false" and "debunked" or "discredited" by the fact check (see Appendix).

Participants were required to spend 15 seconds on the feed before the continue button appeared. After viewing the feed, participants answered a series of post-test questions before being debriefed regarding the purpose of the study. Participants who failed a post-test attention check were excluded from analysis (N=139).

Measures

Story veracity. A single item asked participants to rate the extent to which the story was definitely false – definitely true on a five-point scale, mimicking previous measurement (e.g., Amazeen et al., 2018). A sixth option allowed participants to select "unsure," (n=82) which was combined with the "mixed (both true and false)" option to create a five-point scale (although we also ran the analyses excluding these participants, see footnote 9; M=2.50, $S.D.$=.89).

Issue attitudes. A single question, parallel across experimental conditions, measured issue attitudes. Participants either rated their agreement on seven-point scales from strongly disagree to strongly agree with the statement "Zombie ticks can cause red meat allergies" (M=3.45, $S.D.$=1.24) or "Fluoride in drinking water reduces intelligence" (M=3.01, $S.$

D.=1.36). These items were coded so that a higher score indicates greater agreement with these statements – in other words, more alignment with the argument made in the news story regardless of whether that story is true (ticks) or false (fluoride). Therefore, participants with higher issue attitudes in the ticks conditions are more informed, whereas those with higher issue attitudes in the fluoride conditions are more misinformed.

Behavioral Intentions. Additionally, participants answered one question regarding their behavioral intentions: "I will avoid traveling to areas with zombie ticks" (seven-point scale, *M*=4.01, *S. D.*=1.71) or "I will try to avoid drinking water that has fluoride in it" (*M*=3.61, *S. D.*=1.67). These items are coded so that a higher score involves more avoidance behaviors – in other words, behavioral intentions in alignment with story position.

Testing Recall

We first test RQ1 and RQ2, which asked about recall for the original poster's position (RQ1) and the replies' position (RQ2). We use a series of four chi-squared tests, run separately for the position of the original poster and the reply posts and for the two issues: ticks (true) and fluoride (false).

Per RQ1, we observe two patterns for recall of the original poster's position (Table 2). First, we find a significant effect of our manipulation for both ticks, χ^2 (3,337)=44.10, p<.001, and fluoride, χ^2 (3, 345)=52.56, p<.001, with participants more likely to report the poster said it was true when the poster claimed it was true, and more likely to say the poster said it was false when the poster claimed it was false. However, overall participants were relatively *inaccurate* in remembering the position of the original poster: just over 50% could accurately report that the original poster said the content was "true" (when that is what they posted), whereas roughly 30% could accurately report the poster said the content was "false" (when the poster said that). It

appears that participants default to believing the original poster said the post was true: even when the poster said the post was *false*, roughly a third of participants reported the poster said it was true. This process functions similarly when the information is true (ticks) or false (fluoride).

When considering the recalled position of the reply tweets, per RQ2, we see a significant impact of reply position for both ticks, χ^2 (3,337)=56.73, p<.001, and fluoride, χ^2 (3,346)=31.00, p<.001. We again find that participants were more likely to report the replies said the story was true when the replies claimed it was true, and more likely to say the replies said the story was false when the replies claimed it was false. Recall of the reply tweet position remains low (under 50% for both issues and positions), but in this case, the default assumption for the reply position does not appear to be that the replies say the story is "true" – more participants instead admit they don't know.

Overall across issues and contexts, 42.8% of participants correctly identified the original poster's position, 41.2% correctly identified the position of the reply posts, and only 22.9% of participants correctly identified both the original poster and reply positions. The low recall for these posts is worth keeping in mind as we move to our tests of experimental effects.

Results

To test our hypotheses and research questions regarding the effects of our manipulations on perceived veracity, issue attitudes, and behavioral intentions, we use a series of three-way ANCOVAs, entering original poster position, reply position, and article accuracy as factors, and controlling for whether participants had seen the story before and when they participated in the

study.[7] Notably, in these analyses we do not eliminate people who did not accurately recall the position of the poster or the replies from these analyses – a question we explore later.

Story Veracity

We begin by testing the effects of different truth signals on perceptions of story veracity. We find no support for H1a, as perceptions of story veracity are not impacted by the original poster's position (see Table 4 for all F-values and significance levels). H2a is supported, such that the story is seen as lower in veracity when the replies say it is false than when the replies say the story is true (see Figure 1). We do not observe an interaction between the original poster's position and the reply posters' position, failing to support H3a (and per RQ3a), suggesting original poster position and reply position operate separately for perceptions of story veracity, with reply position mattering more.

Finally, we test whether the effects of correction (either from the original poster or the replies) differ depending on whether the story is actually true (ticks) or false (fluoride). First, we note that there are no overall differences in perceptions of story veracity by article accuracy, reinforcing the results of the pre-test that the two stories were seen as equally (un)true.[8] Per RQ4, our results suggested that story position did not moderate the effects of reply position or the (null) effects of original poster position, nor is there a three-way interaction among original poster position, reply position, and article accuracy. These results suggest that replies function to "debunk" accurate stories as well as inaccurate ones.[9]

Issue Attitudes

---

[7] We additionally test whether story exposure changed over time: Two independent-samples t-tests (performed separately for each story) found that *when* people took the survey did not significantly differ among those who had and had not seen the story before, suggesting exposure did not change throughout the data collection period.

[8] Perceptions of story veracity are low for both the tick (M=2.52, $S.E.$=.05) and fluoride issue (M=2.48, $S.E.$=.05), in line with our pre-test suggesting both headlines were seen as likely to be false.

[9] These effects are consistent when the participants who answered "don't know" for story veracity are dropped from the analyses (N=82) rather than combined into the "mixed" category.

Next, we examine the effects of our experimental manipulations on issue attitudes – namely, whether attitudes align with the position of the shared story. These results largely mimic those for story veracity. Original poster position did not impact issue attitudes, in contrast to H1b, but as predicted by H2b, when the replies say the story is false, individuals report lower levels of issue agreement with the position endorsed by the story (see Figure 1). There is again no interaction between original poster position and reply position in affecting issue attitudes, in contrast to H3b and RQ3b.

Finally, we turn to the role of article accuracy. Here, we observe a main effect of issue, with people in the ticks condition reporting higher levels of issue agreement (M=3.45, $S.E.$=.07) than those in the fluoride condition (M=3.01, $S.D.$=.07). In other words, people's issues were less aligned (and therefore more accurate) with the false (fluoride) story than the true (ticks) story – although it is worth noting that in *both* cases they disagree with the premise of the story (which is accurate for fluoride but a misperception in the case of ticks). However, as before article accuracy does not interact with original poster position, with reply position, or with both in combination. Thus, replies are equally likely to shift an individual's issue attitudes regardless of whether the story is actually true or false.

Behavioral Intentions

Our final analyses consider whether the original poster's position, reply position, and article accuracy influence behavioral intentions for the issue – notably, whether people will try to avoid exposure to the threat raised in the story. These results again resemble the results for story veracity and issue attitudes. Original poster position does not affect issue behavioral intentions, in contrast to H1c. H2c is supported: there is a main effect of reply position, with people reporting more intentions to avoid the threat when the replies say the story is true as compared to

when they say the story is false (see Figure 1). Finally, there is no interaction between original poster position and reply position, suggesting the effects of reply position do not depend on what the original poster said, providing no support for H3c or RQ3c.

Turning to the role of article accuracy, we again find a main effect of article accuracy on behavioral intentions. People report higher intentions to avoid traveling to areas with Zombie ticks (M=4.01, $S.E.$=.09) than to avoid drinking fluoridated water (M=3.61, $S.E.$=.09). Notably, article accuracy again does not interact with original poster position, reply position, or as a three-way interaction among all three factors. As before, article accuracy has no bearing on the ability of the replies, nor the inability of the original poster, to alter behavioral intentions.

Supplemental Analysis

However, as noted above, many people could not recall the position of the original poster or the reply posts even immediately after exposure. Therefore, we replicate these analyses among those with perfect recall for the original poster's position and the reply positions (reducing the sample from 682 to 156). This supplemental analysis supports the role of replies that claim the story is false in reducing perceived story veracity, issue attitudes, and behavioral intentions (see Appendix[10]). The role of reply position remains consistent across both the true and false issues and does not depend on the original poster's position. There is one unexpected main effect of original poster, wherein perceptions of story veracity are *higher* when the original poster says the story is false (M=2.62, $S.E.$=.13) than when they say the story is true (M=2.30, $S.E.$=.09). However, given the small sample size, we consider these results with caution.

Discussion

[10] Supplemental appendix available at: https://doi.org/10.7910/DVN/9Y0SSL

This study makes three main contributions to the literature. First, individuals cannot effectively debunk health misinformation by sharing an inaccurate story and explicitly stating it is false. Second, replies are more effective than an original poster in debunking (or confirming) perceptions of story veracity. Third, "corrections" from replies to a posted story can change perceptions of story veracity, issue attitudes, and behavioral intentions *regardless* of whether the story is actually true or false. We elaborate on the implications of our results for individuals and groups interested in correcting health misinformation on social media below.

First, we find that sharing misinformation on social media in order to debunk it – something nearly 5% of Americans have reported doing (Mitchell, et al., 2019) – is not an effective strategy. The claim of the original poster as to whether a story is true or false does not affect perceptions of story veracity, issue attitudes, or behavioral intentions. This differs from previous research wherein an original poster's credibility predicts news trust or source credibility (Oeldorf-Hirsch & deVoss, 2020; Turcotte et al., 2015). However, this study explicitly compares when the original poster says the story is *false* versus *true,* rather than applying the poster's credibility (held constant) to the material.

Our recall measures provide one explanation for these null effects of the original poster. Only 43% of participants correctly recalled the position of the original poster immediately after exposure, and people appear to assume that the original poster believes the story is true, regardless of what the poster actually said. Sharing misinformation to debunk it should be discouraged, as people are likely to only remember the story itself, not the pre-bunking efforts of the poster. This recommendation aligns with concerns about creating a familiarity effect, by which misinformation seems more credible because it is familiar (Lewandowsky et al., 2012).

Second, we find that when two individuals *reply* to the story saying the information shared in the story is false and providing links to reinforce their claim, people adjust their perceptions of story veracity, issue attitudes, and behavioral intentions to align with these replies. This echoes previous work regarding the power of corrections from other people on social media (Bode & Vraga, 2018; Margolin et al., 2018; Vraga et al., 2020). There are several reasons why the replies may have a larger impact than the original poster. First, people do not default to believing that the replies say the story is "true," as they do for the original poster. Thus, while recall for the reply position remains low (still roughly 41% recall), it is not systematically biased towards endorsement. Second, there are two replies, and both provide links to expert sources to substantiate their claim regarding the article, following best practices for social media correction (Vraga & Bode, 2020) and the number of sources may encourage more thoughtful elaboration and thus persuasion (Lee & Shin, 2021). As such, these multiple replies may be more convincing than a single rebuttal without evidence from the original poster.

Third, this ability of replies to persuade individuals is not entirely good, as this process functions similarly when the story is false (fluoridation does not cause reduced IQ) as when the story is true (Zombie ticks can cause red meat allergies). In other words, people adjust their attitudes in line with position endorsed by the replies, regardless of the accuracy of these replies. Notably, these replies were effective in debunking a story for which the other "truth signals" from the story itself – a low credibility source and stories seen as quite implausible – both lean towards falsity. Future research should test whether replies "debunking" a true story are equally effective when the story has more cues signaling its veracity. Additionally, our study follows best practices by having multiple people respond to "debunk" the story using an expert source, whether the story is actually true or false. Importantly, individuals were able to do so as the

headlines displayed for these links are inconclusive, often using a question. This is not uncommon for fact checking websites (Tsfati et al., 2020) – although one that professional fact checkers increasingly discourage (Gyenes, 2020). Imprecision in headlines enables reply posters to frame the fact check in whatever manner they choose – for example, by claiming that a fact check says a story is "false" when the fact check says it is true (Darcy, 2021). Such choices may be either deliberate to create disinformation, or inadvertent spreading of misinformation – but both are forms of "pseudo-information" with high social consequences (Kim & gil de Zuniga, 2020). Whatever the motive, we suspect such framing is effective, given that most social media links are never clicked on (Gabielkov, Ramachadran, Chaintreau, Legout, 2016). Fact-checking and expert organizations can limit this potential for dis- or mis-information by making the *answer* to the fact check appear in the headline, rather than leaving it open-ended for audiences to reframe as they choose (Stroud, et al, nd).

We acknowledge several limitations of this study. First, we rely upon a young, educated student sample to test these effects, which may impact the response to misinformation and its correction in several ways. One meta-analysis of correction of health misinformation on social media found stronger effects among adult samples, as well as those that are highly involved (Walter et al., 2020). It is possible that a sample more involved in the topics studied may pay more attention to the position of the original poster, for example, unlike those in our sample. Older users of social media may also have different habits – they may use different platforms, use them differently, or have different expectations of behaviors (Pew, 2019) – that could affect what they pay attention to and how information affects them. Older and younger people also react differently to health communications messages in general: for instance, younger populations are more persuaded by repetition (Keller & Lehman, 2008).

Second, we use a simulated feed to test these mechanisms. While pains were taken to make the feed appear as realistic as possible, participants undoubtedly engage differently with content on their own social media feed, and we cannot test that here. We might expect, for example, that when a story is shared by a known peer or social connection, more attention may be paid to their position than when it comes from an unknown user. Future research should triangulate methods to analyze these effects within social media (i.e., Margolin et al., 2018) or by qualitatively exploring people's reactions to these types of posts.

Third, we use stories that are always seen as relatively implausible – even when one story is true. These processes may differ if the stories enjoy higher (or lower) initial credibility. We also focus on relatively unknown health claims unlikely to provoke partisan beliefs, rather than the health topics like tobacco, vaccination, and epidemics commonly studied (Krishna & Thompson, 2021), which may produce different outcomes given their societal prominence. We test our hypotheses using cross-sectional data, which does not allow us to speak to the endurance of any misperceptions that may arise (or be mitigated) in our study.

Finally, we created corrections that relied on recognizable expert sources with limited perceived biases – in this case, Snopes.com and USA Today. While meta-analyses demonstrate that corrections from experts are more successful as compared to non-experts (Walter et al., 2020), more research is needed regarding different sources of correction and the signals of expertise incorporated. Additionally, identifying sources that are seen as trustworthy and expert across societal groups is increasingly difficult; it behooves researchers to consider which sources of correction are most effective for particular issues, groups, or platforms.

Together, this study offers a number of practical and theoretical takeaways for researchers, physicians and other health communication specialists, people and organizations

using social media, social media platforms, and fact checking organizations. First, social media users, as well as health communication specialists, should know that sharing misinformation in order to debunk it is not an effective strategy, as it can reinforce the misinformation rather than correcting it. When attempting to prebunk misinformation, individuals and groups should focus on promoting highly credible information from experts on the topic or highlighting corrective messages, rather than naming and linking to specific misinformation (Lewandowsky et al., 2020).

Responding to misinformation, however, remains a powerful mechanism to reduce perceptions of story veracity and misperceptions, and should be encouraged among social media audiences. They should also know that this process functions similarly for stories that are both true and false – at least those that *seem* to be false and are later "verified" as false by others' responses. This suggests caution and verification when encountering replies that claim to debunk misinformation. Efforts to automatically detect deceptive comments may be an important tool for researchers and social media companies to identify and address these misleading responses (e.g., Oh & Park, 2021). Social media companies can make use of these findings to tweak the design of their platforms, informing users that sharing misinformation to debunk it is ill advised, encouraging them to debunk misinformation by replying to it, and boosting the visibility of those replies which do so. Fact checking organizations should ensure that their headlines include a clear determination of the veracity of misinformation, to prevent audiences from leveraging their credibility to support a false conclusion. Likewise, health professionals creating shareable content for social media should prioritize engaging responses that clearly state the best evidence and scientific consensus on the issue within the graphic or post. Health misinformation continues

to present challenges for audiences to interpret, but individuals and platforms can respond in ways that limit its negative effects.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211-36.

Avaaz. (2020). *Facebook's algorithm: A major threat to public health.* Retrieved from: https://secure.avaaz.org/campaign/en/facebook_threat_health/

Bigsby, E., & Hovick, S. R. (2018). Understanding associations between information seeking and scanning and health risk behaviors: An early test of the structural influence model. *Health Communication*, *33*(3), 315-325.

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*, 619–638. DOI: 10.1111/jcom.12166

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication*, *33*(9), 1131-1140.

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annual Review of Political Science*, *10*, 103-126.

Darcy, O. (2021, March 18). Headlines lacking context exploited by anti-vaccine activists to wrongly suggest danger, study finds. *CNN.com.* Retrieved from: https://www.cnn.com/2021/03/18/media/covid-19-vaccine-headlines/index.html

Druckman, J. N. (2004). Political preference formation: Competition, deliberation, and the (ir) relevance of framing effects. *American Political Science Review*, 671-686.

Ecker, U. K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*(3), 570-578.

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016, June). 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* (pp. 179-192).

Gyenes, N. (2020). A standard of care for health misinformation: lessons from Global Fact. Retrieved from: https://meedan.com/blog/a-standard-of-care-for-health-misinformation-lessons-from-global-fact/

Harris, R. (2019). Can Maternal Fluoride Consumption During Pregnancy Lower Children's Intelligence? *NPR.* https://www.npr.org/sections/health-shots/2019/08/19/752376080/can-maternal-fluoride-consumption-during-pregnancy-lower-childrens-intelligence

Jolley, D., & Douglas, K. M. (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PloS one*, *9*(2), e89177.

Kang, H., Bae, K., Zhang, S., & Sundar, S. S. (2011). Source cues in online news: Is the proximate source more powerful than distal sources?. *Journalism & Mass Communication Quarterly*, *88*(4), 719-736.

Keller, P. A., & Lehmann, D. R. (2008). Designing effective health communications: a meta-analysis. *Journal of Public Policy & Marketing*, *27*(2), 117-130.

Knight/Gallup. (2018, June 20). Americans' views of misinformation and how to counteract it. *Knight Foundation.* Retrieved from: https://knightfoundation.org/reports/americans-views-of-misinformation-in-the-news-and-how-to-counteract-it/

Kim, J. N., & de Zúñiga, H. G. (2021). Pseudo-Information, Media, Publics, and the Failing Marketplace of Ideas: Theory. *American Behavioral Scientist*, *65*(2), 163-179.

Krishna, A., & Thompson, T. L. (2021). Misinformation about health: A review of health communication and misinformation scholarship. *American Behavioral Scientist, 65*(2), 316-332.

Lee, E., & Shin, S. Y. (2021). Mediated misinformation: Questions answered, more questions to ask. *American Behavioral Scientist, 65*(2), 259-276.

Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., et al.. (2020). The Debunking Handbook 2020. Available at https://sks.to/db2020. DOI:10.17910/b7.1182

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.

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political Fact-Checking on Twitter: When Do Corrections Have an Effect? *Political Communication, 35,* 196-219.

Metaxas, P. T., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., & Finn, S. (2015, April). What Do Retweets Indicate? Results from User Survey and Meta-Review of Research. In *ICWSM* (pp. 658-661).

Molina, M. D., Sundar, S. S., Le, T., & Lee, Dongwon. (2021). "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist, 65*(2)*,* 180-212.

Molyneux, L., & Mourão, R. R. (2019). Political journalists' normalization of Twitter: Interaction and new affordances. *Journalism Studies*, *20*(2), 248-266.

Oeldorf-Hirsch, A., & DeVoss, C. L. (2020). Who posted that story? Processing layered sources in Facebook news posts. *Journalism & Mass Communication Quarterly*, *97*(1), 141-160.

Oh, Y. W., & Park, C. H. (2021). Machine cleaning of online opinion spam: Developing a machine-learning algorithm for detecting deceptive comments. *American Behavioral Scientist, 65*(2), 389-403.

Mitchell, A., Gottfried, J., Stocking, G., Walker, M., & Fedeli, S. (2019). Many Americans Say Made Up News Is a Critical Problem That Needs To Be Fixed. Pew Research Center. Retrieved from: https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/

Pew. (2019, June 12). Social media fact sheet. *Pew Research Center.* Retrieved from: https://www.pewresearch.org/internet/fact-sheet/social-media/

Porter, E., & Wood, T. J. (2019). *False Alarm: The Truth About Political Mistruths in the Trump Era*. Cambridge University Press.

Smith, C. N., & Seitz, H. H. (2019). Correcting Misinformation About Neuroscience via Social Media. *Science Communication*, *41*(6), 790-819.

Sülflow, M., Schäfer, S., & Winter, S. (2019). Selective attention in the news feed: An eye-tracking study on the perception and selection of political news posts on Facebook. *new media & society*, *21*(1), 168-190.

Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news" A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137-153.

Tandoc Jr, E. C., Lim, D., & Ling, R. (2020). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, *21*(3), 381-398.

Tsfati, Y., Boomgaarden, H. G., Strömbäck, J., Vliegenthart, R., Damstra, A., & Lindgren, E. (2020). Causes and consequences of mainstream media dissemination of fake news:

literature review and synthesis. *Annals of the International Communication Association*, *44*(2), 157-173.

Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication*, *20*(5), 520-535.

van der Meer, T. G., & Jin, Y. (2020). Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health Communication*, *35*(5), 560-575.

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication, 39*(5), 621-645.

Vraga, E. K., & Bode, L. (2020). Correction as a solution for health misinformation on social media. *American Journal of Public Health, 110*(S3), S278-S280.

Vraga, E., Bode, L., & Troller-Renfree, S. (2016). Beyond self-reports: Using eye tracking to measure topic and style differences in attention to social media content. *Communication Methods and Measures*, *10*(2-3), 149-164.

Vraga, E. K., Kim, S. C., Cook, J., & Bode, L. (2020). Testing the effectiveness of correction placement and type on Instagram. *The International Journal of Press/Politics*.

Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2020). Evaluating the impact of attempts to correct health misinformation on social media: a meta-analysis. *Health Communication*, 1-9.

Wardle, C. (2017, Feb. 16). Fake news. It's complicated. *Medium.com.* Retrieved from: https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79

Wilner, T., & Holton, A. (2020). Breast Cancer Prevention and Treatment: Misinformation on Pinterest, 2018. *American Journal of Public Health*, *110*(S3), S300-S304.

Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Tables and Figures

Table 1: Pre-Test evaluations of headlines (N=100)

Headline	Seen	Veracity	Certainty	Plausible	Not believable	Likely false
Ticks	23%	2.32	2.72	2.46	3.67	3.70
Fluoride in water	24%	2.16	2.69	2.44	3.59	3.63

Table 2: Recall for the position of the original poster

		Zombie Ticks (n=337)		Fluoride (n=345)	
Actual position of Original Poster		False	True	False	True
Participants say:	True	37.4%	**56.4%**	32.4%	**52.7%**
	False	**27.6%**	2.5%	**35.8%**	5.3%
	Uncertain	12.6%	20.2%	5.7%	13.6%
	Don't know	22.4%	20.9%	26.1%	28.4%

Bolded when participants are accurate.

Table 3: Recall for the position of the reply posts

		Zombie Ticks		Fluoridation	
Actual position of Replies		False	True	False	True
Participants say:	True	16.4%	**47.6%**	14.1%	**32.6%**
	False	**40.9%**	10.2%	**44.1%**	19.4%
	Uncertain	14.0%	12.7%	7.1%	11.4%
	Don't know	28.7%	29.5%	34.7%	36.6%

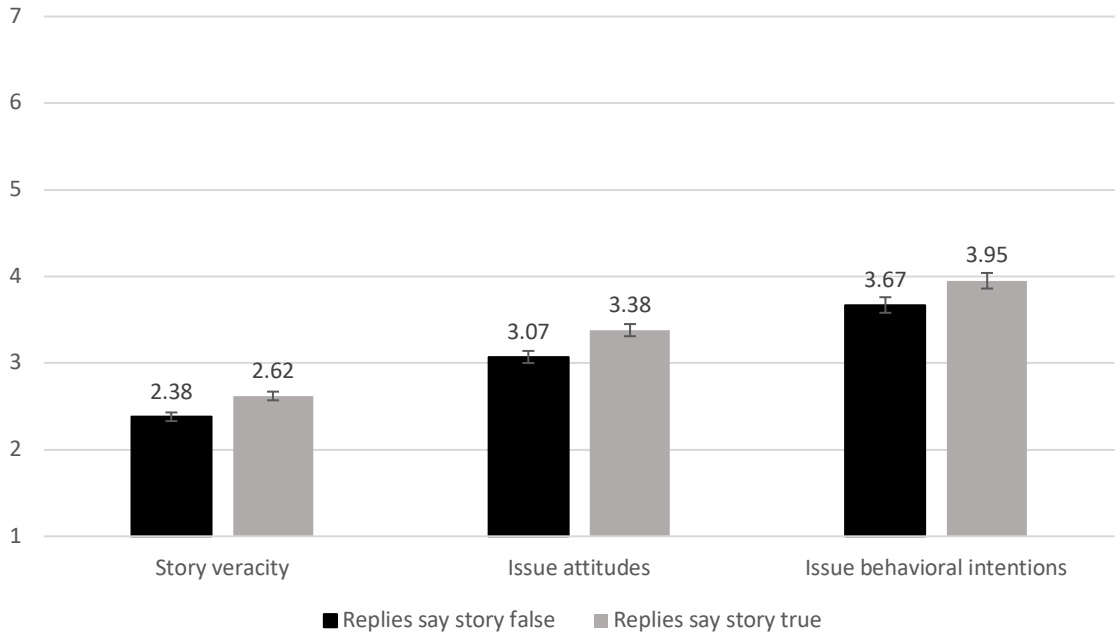
Bolded when participants are accurate.

Table 4: Experimental results using three-way ANCOVAs

	Full sample (N=682)		
	F-value	P-value	η_p^2
Veracity, including those "unsure" with "mixed" at the scale mid-point			
When survey taken	.34	.56	.001
Seen story before	19.78	>.001	.029
Original Poster Position	.00	.99	.00
Article Accuracy	.41	.52	.001
Reply Poster Position	**13.18**	**>.001**	**.019**
Article Accuracy * Reply Position	.01	.92	.000
Article Accuracy * Original Poster	1.31	.25	.002
Original Position * Reply Position	2.42	.12	.004
Article Accuracy * Original Position * Reply Position	.38	.54	.001
Veracity (dropping those who said they were "unsure")			
When survey taken	.46	.50	.001
Seen story before	26.77	>.001	.043
Original Poster Position	.02	.90	.00
Article Accuracy	.47	.49	.001
Reply Poster Position	**15.66**	**>.001**	**.026**
Article Accuracy * Reply Position	.05	.82	.000
Article Accuracy * Original Poster	1.94	.16	.003
Original Position * Reply Position	1.34	.25	.002
Article Accuracy * Original Position * Reply Position	.49	.49	.001
Issue Attitude			
When survey taken	.00	.95	.000
Seen story before	3.91	.05	.006
Original Poster Position	.91	.34	.001
Article Accuracy	**20.53**	**>.001**	**.030**
Reply Poster Position	**9.25**	**>.01**	**.014**
Article Accuracy * Reply Position	.59	.44	.001
Article Accuracy * Original Poster	2.16	.14	.003
Original Position * Reply Position	.30	.58	.000
Article Accuracy * Original Position * Reply Position	.03	.86	.000
Behavioral Intentions			
When survey taken	1.13	.29	.002
Seen story before	.13	.72	.000
Original Poster Position	1.37	.24	.002
Article Accuracy	**9.62**	**>.001**	**.014**
Reply Poster Position	**4.74**	**.03**	**.007**
Article Accuracy * Reply Position	1.18	.28	.002
Article Accuracy * Original Poster	3.67	.06	.005
Original Position * Reply Position	2.29	.13	.003
Article Accuracy * Original Position * Reply Position	1.52	.22	.002

Significant effects reported in bold

Figure 1: Effects of reply position on story veracity and issue attitudes (N=682)



NOTE: Story veracity is measured on a 5-point scale; issue attitudes and behavioral intentions are measured on a seven-point scale. Higher numbers indicate higher veracity and more agreement in terms of attitudes and behavioral intentions with story position. As a result, higher story veracity, issue attitudes, and behavioral intentions represent *accuracy* for the ticks issue and *misperceptions* for the fluoride issue.