Cite as: Vraga, E. K., Kim, S. C., & Cook, J. (2019). Testing Logic-based and Humor-based Corrections for Science, Health, and Political Misinformation on Social Media. *Journal of Broadcasting & Electronic Media*, *63*(3), 393-414.

Testing logic-based and humor-based corrections for science, health, and political misinformation on social media

By:

Emily K. Vraga[1]

Sojung Claire Kim[2]

John Cook[3]

Keywords: misinformation, inoculation, correction, social media

---

[1] Emily Vraga is an associate professor in the Hubbard School of Journalism and Mass Communication at the University of Minnesota. She can be reached at 338 Murphy Hall, 206 Church St SE, Minneapolis, MN 55455. Her email is ekvraga@umn.edu. She is the corresponding author on this paper.

[2] Sojung Claire Kim is an assistant professor in the Department of Communication at George Mason University. She can be reached at 121 NE Module, MS 3D6, 4400 University Dr., Fairfax, VA 22030. Her email is skim205@gmu.edu

[3] John Cook is a research assistant professor in the Center for Climate Change Communication at George Mason University. He can be reached at 247 Research Hall, MS 3D6, 4400 University Dr., Fairfax, VA 22030. His email is jcook20@gmu.edu

Abstract

Misinformation causes a range of negative impacts. One proposed solution is applying critical thinking techniques to neutralize misinformation by explaining its misleading techniques or logical fallacies. This study tests the efficacy of corrections after exposure to misinformation that adopt inoculating techniques. We test two forms of correction—logic-based and humor-based—across the issues of climate change, gun control, and HPV vaccination. We find that results vary across topics, with both logic-based and humor-based corrections reducing misperceptions only for HPV vaccination. More research is needed to test the efficacy of logic-based and humor-based corrections across different issues.

Keywords: misinformation, inoculation, correction, social media

Testing logic-based and humor-based corrections for science, health, and political misinformation on social media

Misinformation can cause a range of negative impacts on democratic societies, which depend on a well-informed populace to properly function. Misinformation can lead to public preferences different to if they were accurately informed, which can have negative policy implications (Kuklinski, Quirk, Jerit, Schwieder, and Rich, 2000) and lead to suboptimal health behaviors (Tan, Lee, & Chae, 2015). Not only can misinformation cause people to believe false information, it can cancel out the impact of accurate information and lead people to disengage from facts (Cook, Lewandowsky, & Ecker, 2017; McCright, Charters, Dentzman, and Dietz, 2016; van der Linden, Leiserowitz, Rosenthal, & Maibach, 2017).

Many scholars have expressed concerns that social media have exacerbated the influence of misinformation, documenting the prevalence of misinformation on a range of topics across social media platforms, with special emphasis on Twitter (Broniatowski et al., 2018; del Vicario et al., 2017; Shao et al., 2018). Consequently, a body of research has grown studying how to counter the influence of misinformation (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). Much of the research focuses on corrective responses after misinformation has been encountered. Refutation is difficult, as misinformation continues to influence people even after reading retractions (Seifert, 2002; Thorson, 2016). However, some techniques increase the effectiveness of refutations. Examples of refutational best-practices include providing factual alternatives to replace the debunked myth (Ecker, Lewandowsky, Cheung, & Maybery, 2015) or providing an explicit warning that recipients are about to be exposed to misinformation to place them on cognitive alert (Ecker, Lewandowsky, & Tang, 2010; but see Clayton et al., 2019).

Even as many decry the role of social media in spreading misinformation, social media offer unique opportunities to immediately correct misinformation. Despite the potential for echo chambers to form, heterogeneous relationships characterized by weak ties still exist on many social media platforms (Kim, Chen, & gil de Zuniga, 2013; Kim & Chen, 2016). As a result, people are likely to come across those who disagree with their views, creating opportunities for correction to occur. These corrections can be effective, reducing misperceptions for individuals sharing misinformation (Margolin, Biones, & Waks, 2018), as well as for the community seeing the interaction (Bode & Vraga, 2015; Vraga & Bode, 2017; 2018). It is this latter process, termed *observational correction* (Vraga & Bode, 2017, p. 14), that we explore in this paper.

Previous research has examined corrections that directly refuted the incorrect information, providing links to expert sources (e.g., Margolin et al., 2018; Vraga & Bode, 2017). However, other studies have suggested that inoculation theory, which describes the techniques of misinformation and present counter-arguments to undermine the logical fallacies, may also be an effective way to neutralize misinformation (McGuire & Papageorgis, 1961).

This study combines these two approaches to addressing misinformation on social media: inoculation and observational correction. We use the theoretical premise of inoculation theory – that describing the techniques of misinformation and exposing its flaws undermines its effectiveness – in correcting misinformation shared in a Twitter post about three controversial issues: climate change, gun control, and the HPV vaccination. This approach of exposing recipients to a "weak form of misinformation" after exposure to the original misinformation has been described as therapeutic inoculation (Compton, 2017). We further explore whether a logic-based versus a humor-based explanation of the fallacious reasoning in the misinformation would affect credibility ratings for the inaccurate post and reduce misperceptions among those seeing

the exchange. This study offers insight into the promise of therapeutic inoculation as a correction strategy and whether its effects differ by correction approach and issue context.

Inoculation theory as a strategy to correct misinformation

Inoculation messages generally contain two structural elements—a forewarning of the danger of being misinformed and counter-arguments refuting the misinformation. The forewarning is designed to elicit threat in recipients, generating a motivation to protect positions now perceived to be vulnerable (Banas & Richards, 2017). The counter-arguments elicit refutational preemption and models counterarguing for recipients.

Inoculation has been shown to be effective in reducing misperceptions across a range of issues, including health (Compton, Jackson, & Dimmock, 2016), politics (Compton & Ivanov, 2013), and climate change (Cook et al., 2017; van der Linden et al., 2017), as well as across modes and audiences (Banas & Rains, 2010).

Pre-emptive refutation before recipients are exposed to misinformation has been shown to be more effective at reducing the influence of misinformation than reactive refutation (Bolsen & Druckman, 2015). Just as exposure to a weak form of a virus conveys resistance to recipients prior to exposure to a real virus, similarly, exposure to weak forms of misinformation reduces the influence of subsequently encountered misinformation (McGuire & Papageorgis, 1961). But given that misinformation techniques are constantly evolving (e.g., Allcott, Gentzkow, & Yu, 2018; Bronianowski et al., 2018; Pew, 2017), pre-emptive inoculation may not always be available across a range of issues and domains, making it important to also test reactive methods for constructing corrective messages.

Inoculating messages may take a range of different forms. Most inoculation research has tested fact-based inoculations, where information about a topic is used to counter misinformation

on the same topic. However, recent research has begun to explore logic-based inoculations, which involve critical thinking approaches of explaining the fallacious reasoning in misinforming arguments (Cook et al., 2017). This harkens back to some of the earliest forms of inoculation proposed by Aristotle, who argued that understanding reasoning flaws offered a universal safeguard against misinformation (Compton, 2005). This critical thinking approach also echoes goals from media and news literacy efforts (Mihailidis & Viotty, 2017; Tully, Vraga, & Bode, 2019). Highlighting the flaws in specific misinformation content after exposure should undermine its credibility and its impact on attitudes. Consequently, efforts to develop critical thinking strategies and resources documenting reasoning fallacies in misinformation are ongoing (Cook, Ellerton, & Kinkead, 2018).

However, warnings in response to online misinformation carry risks. Labelling Facebook misinformation as "fake news" led to increased sharing of the article and a surge in traffic (Levin, 2017), and can create an "implied truth" effect for any stories *not* labeled, even when inaccurate (Pennycook & Rand, 2017). Similarly, Facebook users with conspiratorial leanings increased their engagement with conspiratorial posts after encountering corrections (Zollo et al., 2017). Therefore, it is important that social media responses to misinformation are informed by findings from social science research—an approach known as technocognition (Lewandowsky et al., 2017). This involves an interdisciplinary combination of research findings from communication, psychology, critical thinking, behavioral economics, and computer science, in designing scalable, technological solutions.

Using humor to communicate science and critical thinking

Parallel argumentation offers an alternative method to disseminate logic-based interventions. Parallel arguments adopt the same logical structure as a misinforming argument,

but applied in an absurd situation to clearly communicate the false logic. This approach offers strong pedagogical value as it makes use of concrete examples closer to everyday experience than abstract reasoning (Juthe, 2009). Parallel arguments are also conducive to being delivered in humorous form, and humor is known to produce unique benefits for science communication.

First, humor acts as a gateway that can introduce audiences to scientific information—engaging audiences who are otherwise disengaged from an issue (e.g., Becker & Waisman, 2013). For example, satirical messages have been found to raise acceptance of climate change particularly amongst people who have low interest in the issue (Brewer & McKnight, 2015; 2017), or increase young people's intentions to engage in activism (Bore & Reid, 2014; Skurka, Niederdeppe, Romero-Canyas, Acup, 2018).

Second, humor can impact how people process the message. Humor can boost a messenger's credibility, especially when the target is unknown (Vraga, Johnson, Carr, Bode, & Bard, 2014), but discourage active scrutiny of the message, signaling that critical thought is unnecessary (Young, 2008). Peripheral processing may explain competing findings about the effects of humor on learning, depending on the issue and personal predispositions (e.g., Becker & Bode, 2018; Becker & Waisman, 2013; Kim & Vishak, 2008).

Visual humor (e.g., comics) are one mode of humorous communication that offers further advantages. Visual memes are particularly relevant and effective for younger audiences, who largely use image-based social media platforms such as Instagram, Snapchat, and YouTube (Pew Research Center, 2018). Scientific visualizations in comic form can improve knowledge acquisition and problem-solving skills (Farinella, 2018). The way comics integrate text and images can improve learning, especially for complex topics—known as the spatial contiguity

effect (Jee & Anggoro, 2012). Comics are also conducive to analogy, making abstract topics more concrete and accessible to readers.

This study tests whether two inoculation approaches—logic-based versus humor-based in cartoon form—that expose the flaws in misinformation reasoning can work in a corrective response across three controversial issues: climate change, gun control, and HPV vaccination. These two approaches were tested in the form of a response to a misinforming tweet.

Correction on Social Media

Social media provide an important test for whether inoculation strategies included in corrective messages reduce misperceptions. Social media have been criticized for facilitating the spread of misinformation on a range of political (Lewandowsky et al., 2017; Tucker et al., 2018), health (Bronianowski et al., 2018; Guidry et al., 2015), and scientific topics (Bessi et al., 2015; del Vicario et al., 2017).

Despite the propensity for misinformation to spread on social media, recent work into observational correction has suggested that user correction involving multiple users offering links to expert sources can effectively reduce misperceptions among the community (Vraga & Bode, 2017; 2018), and corrections are even stronger when corrections come from known others (Margolin et al., 2017). However, evaluations of these corrective responses are often subject to motivated reasoning depending on participants' predispositions (Bode & Vraga, 2015).

In this study, we expand on observational correction in several ways. First, existing studies have focused on directly rebutting the misinformation using expert sources (e.g., GMO food *is* safe to consume) rather than pointing out the logical fallacies in the misinformation, as examined in inoculation theory. We test whether a single user offering this detailed rebuttal is sufficient to correct misinformation. Second, previous research has focused on health and

scientific issues (e.g., Zika, GMO foods), rather than exploring its effects across a range of issues, including an explicitly political issue like gun control. Third, we explore the effects of these responses on perceptions of the credibility of the misinformation itself, as well as evaluations of the corrective responses. We focus on message credibility, but several traditional news credibility measures – such as completeness or authenticity – may not be appropriate for short social media posts (Appelman & Sundar, 2016; Flanagin & Metzger, 2007). Instead, we adapt previous measures of credibility for user corrections, as they most closely relate to this study (Vraga & Bode, 2018).

Developing Expectations

Building upon the research into therapeutic inoculation and observational correction leads us to examine three related processes in this study: the effects of these corrective responses on perceptions of the *credibility* of the original misinformation tweet, on the *credibility* of the corrective responses, and on the *change in misperceptions* across the three issues examined. If the inoculation-based corrective responses function as expected, pointing out the logical fallacies – either in logic-based or humor-based form – in the original misinformation tweet should reduce credibility assessments of the inaccurate tweet as well as improve accurate attitudes on the issue compared to a control of misinformation absent correction.

H1: Exposure to corrective information will reduce (a) credibility ratings of the misinformation tweet and (b) issue misperceptions as compared to a control condition.

However, we are less certain whether the logic-based versus humor-based corrections will be more effective. If humor promotes message engagement and credibility (Becker, 2018; Vraga et al., 2014), it may be more successful than a logic-based correction. However, a previous study found that humorous fact-checks were no more effective than non-humorous fact-checks in

producing message attention or reducing misperceptions (Young, Jamieson, Poulsen, & Goldring, 2017), and others argue that humor can deter effortful processing (Kim & Vashik, 2009; Young, 2008), leading to mixed findings regarding the effects of humor (Becker & Waisman, 2013). Therefore, we ask:

RQ1a: Will the logic-based or humor-based corrections be more effective in reducing (a) credibility ratings of the misinformation tweet and (b) issue misperceptions?

RQ1c: Will the logic-based or humor-based correction be seen as more credible?

Our design also allows us to consider the conditional effects of the logic-based and humor-based corrections. Previous research into observational correction has found effects tend be stronger among those who have higher pre-existing misperceptions on the issue (Bode & Vraga, 2015; Vraga & Bode, 2017), but other research has found that corrections largely reduce misperceptions among those who already hold more accurate attitudes (Garrett & Weeks, 2013) or found no differences across groups (Nyhan & Reifler, 2015). This discrepancy may result from the issue context and type of correction being examined, so we test whether misperceptions condition the response to the logic-based and humor-based corrections in terms of credibility evaluations and change in misperceptions, as well as consider whether the effects of correction differ depending on the issue context: political, health, and science.

RQ2: Will the process of correction differ depending on the issue context: political, health, or science?

RQ3: Will the effects of logic-based and humor-based corrections on (a) credibility ratings of the misinformation tweet and (b) issue misperceptions vary depending on pre-existing issue misperceptions?

Methods

To test our expectations, we employed an experiment embedded in an online survey. Participants were recruited from Amazon's Mechanical Turk, a crowdsourcing service wherein people are paid to complete short tasks. 406 participants were paid $1.00 for a 15-minute survey in May of 2018. Our median participant was male (54%), white (80%), 25-34 years old, and had an associate's degree.

Our experiment used a 3 (Misinformation only, Humor-based correction, Logic-based correction) X 3 (Topic: Climate change, gun control, HPV vaccination) between-subject experimental design. Participants were asked to review tweets about current events, then saw a series of four tweets. Three of these tweets were dummy tweets consistent across conditions; participants were randomly assigned to view one of the nine experimental conditions.

Our first manipulation altered the presence and type of corrective response to misinformation. For three issues (described below), misinformation on the controversial topic contained a logical flaw. In the control conditions, this misinformation was not corrected. In both correction conditions, the response highlighted the logical fallacy invoked in the misinformation. In the logic-based inoculation, an infographic that visualized logical flow was used to explicitly name the logical fallacy (e.g., jumping to conclusions), whereas in the humor-based inoculation a parallel argument in the form of a humorous cartoon was used to illustrate the flawed logic. See the supplemental appendix for the full language of the experimental tweets.

Our second manipulation was of the issue context. We selected controversial topics across three domains. For science, we tested misinformation arguing that climate change is not man-made. For politics, we assessed misinformation arguing that gun control laws do not reduce crime. For health, we examined misinformation arguing that the HPV vaccine causes auto-

immune disorders. We chose these topics based on their controversial nature, their societal impact, and the prevalence of misinformation on social media (e.g., Harvey et al., 2017; Madden, Nan, Briones, & Waks, 2011).

Across all manipulated tweets, we selected gender-neutral names and generic user icon images that did not contain visual cues to reduce the potential for gender cues to influence response to the misinformation or the corrections.

Measures

Original tweet credibility. Participants rated their perceptions of the credibility of the original tweet using a series of three semantic differentials measured on five-point scales: credible, informative, and accurate, adapted from Vraga and Bode (2018). These measures were combined into an index (*cronbach's* α = .91, *M* = 2.57, *S.D.* = 1.19).

Reply tweet credibility. The same three items were used to evaluate the credibility of the reply tweets, which were combined into an index (*cronbach's* α = .92, *M* = 3.54, *S.E.* = 1.18).

Change in issue beliefs. In the pre-test, participants rated their percentage agreement with a series of statements, including the misinformation on the three manipulated issues: "human activity is causing climate change" (*M* = 75.79, *S.D.* = 29.09), "there is no link between HPV vaccine and auto-immune symptoms" (*M* = 49.48, *S.D.* = 33.68), and "gun control laws reduce crime" (*M* = 54.61, *S.D.* = 36.91). These same statements were asked in the post-test as well (climate change: *M* = 77.67, *S.D.* = 28.24, HPV: *M* = 56.84, *S.D.* = 35.49, gun control: *M* = 56.49, *S.D.* = 35.16), and a change score was computed for each issue (climate change: *M* = .35, *S.D.* = 10.76, HPV: *M* = 8.39, *S.D.* = 30.56, gun control: *M* = 1.81, *S.D.* = 18.89).

Results

Manipulation Checks

First, we examine whether people recalled seeing the manipulated tweet. Immediately after seeing the tweets, participants were asked to report if they had seen a tweet about each experimental topics. Overall, recall for the manipulated tweet was high, with 91.9% reporting seeing the manipulated tweet. However, a logistic regression suggests that recall was impacted both by the topic of the tweet, as well as the presence of a corrective response. Recall for the tweet was lowest for the gun issue (87%), significantly lower than HPV tweet (95%, *odds ratio*=3.37, p=.01) and marginally lower than the climate change tweet (93%, *odds ratio*=2.12, p=.09). Likewise, exposure to either the cartoon reply (93%, *odds ratio*=2.44, p=.04) or the logic reply (97%, *odds ratio*=5.99, p=.00) boosted recall compared to misinformation only (86%), which makes sense as they included more content (original tweet plus reply).

Second, we test whether our manipulation of humor in the corrections was effective. Participants were asked to rate whether the reply tweet was humorous/not humorous and entertaining/not entertaining on five-point semantic differential scales (r=.64, p<.001, M=2.78, *S.D.*=1.19). We perform an ANCOVA with two factors: the three issues and the two reply tweet types (cartoon vs. logic), controlling for pre-test attitudes on the manipulated issue. These analyses excluded those in the misinformation-only conditions, who did not see a reply tweet. The ANCOVA confirms a main effect of reply type, F(1, 262)=8.84, p<.001, *partial* η^2=.033, with the humor-based correction (M=2.97, *S.E.*=.10) being rated as more humorous than the logic-based correction (M=2.55, *S.E.*=.10). Perceptions of the humor of the reply tweet was not impacted by issue domain, F(2, 262)=1.68, p=.19, *partial* η^2=.013, nor was the interaction between correction type and topic significant, F(2, 262)=1.21, p=.30, *partial* η^2=.009.

Effects on credibility

We use an ANCOVA to compare the impact of the corrective responses and the issue

context, controlling for pre-existing attitudes on the manipulated issues. We first test the credibility of the misinformation tweet. We find no effect of exposure to a corrective response in ratings of the credibility of the original post, $F(2, 395)=.34$, $p=.71$, *partial* $\eta^2=.002$, in contrast to H1a, nor do corrections differentially impact credibility across issue domains, $F(4, 395)=1.57$, $p=.18$, *partial* $\eta^2=.016$, per RQ2. Interestingly, there is a main effect of topic, $F(2,395)=3.33$, $p=.04$, *partial* $\eta^2=.017$, with the vaccination misinformation tweet receiving significantly lower credibility ratings ($M=2.35$, *S.E.*$=.10$) than the climate change ($M=2.68$, *S.E.*$=.10$, $p=.02$) and gun control misinformation ($M=2.64$, *S.E.*$=.10$, $p=.04$), using a Least Standard Differences comparison. It appears that with the topic of vaccination, the misinformation tweet was seen as less credible overall than for either the climate change or gun control topic, regardless of the presence of a correction.

We replicate these analyses for perceptions of the credibility of the corrective responses. Evaluations of the corrective responses depended both on the type of response, $F(1, 262)=4.93$, $p=.03$, *partial* $\eta^2=.018$, and the topic, $F(2, 262)=5.17$, $p=.01$, *partial* $\eta^2=.038$. The logic-based correction was rated as significantly more credible ($M=3.70$, *S.E.*$=.10$) than the humor-based correction ($M=3.40$, *S.E.*$=.10$), answering RQ1c. Moreover, per RQ2, both corrective responses were seen as less credible for the climate change topic ($M=3.23$, *S.E.*$=.12$) compared to the gun control ($M=3.63$, *S.E.*$=.11$, $p=.02$) and HPV vaccination ($M=3.79$, *S.E.*$=.13$, $p<.001$) topics. There is no interaction between correction type and topic, $F(2, 262)=.52$, $p=.59$, *partial* $\eta^2=.004$.

Effects on misperceptions

Next, we test whether the corrective responses reduce misperceptions on the topic. We find a marginally significant main effect of correction on misperceptions, $F(2, 395)=2.63$, $p=.07$, *partial* $\eta^2=.013$, in line with H1b. Examining the pairwise comparisons suggest that only the

logic-based correction was effective in increasing accurate issue attitudes across topic (M=7.08, *S.E.*=1.76, p=.03) using the LSD comparison, as compared to the misinformation-only condition (M=1.49, *S.E.*=1.78), with the humor-based correction falling between these extremes (M=3.18, *S.E.*=1.76), per RQ1b. However, this effect is conditioned by issue context, F(4, 395)=3.46, p=.01, *partial* η^2=.034, per RQ2. These results indicate that corrective responses were most effective with the HPV vaccination issue. For the HPV vaccination issue, exposure to the humor-based correction increases accurate attitudes as compared to the control (p=.05), while exposure to the logic-based corrections produces more accurate attitudes as compared to either the control (p<.001) or the humor-based response (p=.04). These differences are not significant for either the gun control or the climate change issue (see Figure 1).

Considering Pre-Existing Attitudes

Previous research into observational correction suggests that its effects are conditioned by an individual's pre-existing attitudes on the issue, with effects largely occurring among those with higher initial misperceptions (Vraga & Bode, 2015). Therefore, to answer RQ3, we test pre-existing beliefs as a moderator of correction type in our model separately for climate change, gun control, and HPV vaccination, using Model 1 in PROCESS 3.1 (Hayes, 2017). We parse significant interactions by comparing those with 10% agreement (dismissive), 50% (undecided), and 90% (convinced) agreement with scientific consensus on the issue in the pre-test.

We first examine perceptions of the credibility of the original tweet. For the climate change and the gun control issue, we again see no significant effect of the corrective response on its own or in conjunction with pre-existing beliefs. However, for the HPV issue, we find a significant higher order interaction, F(2,132)=3.24, p=.04, R^2 change=.038, as well as a significant interaction between the logic-based correction and pre-existing attitudes (b=.02,

S.E.=.01, p=.01). Among the dismissive, exposure to the logic-based correction depressed evaluations of the original tweet as compared to the misinformation only condition (p=.04), as we would expect. However, among the convinced, the original misinformation tweet is rated *more* highly when people see a logic-based correction as compared to the misinformation-only condition (p=.08, see Figure 2).

Second, we explore perceptions of credibility for the response tweet. We find that pre-existing beliefs moderate credibility assessment for the corrective responses for all three topics: climate change, F(1,90)=5.75, p=.02, R^2 change=.051, HPV, F(1,75)=3.72, p=.06, R^2 change=.044, and gun control, F(1,92)=3.14, p=.08, R^2 change=.026. However, the pattern of these effects is not consistent across the issues (see Figure 3). For the gun control issue, we see that the logic-based correction is rated more highly than the humor-based correction among the convinced (p=.05), with no differences among the dismissive (p=.47) and undecided (p=.47). Similarly, for climate change the logic-based approach fares better among the convinced (p=.08), with no difference among the undecided (p=.33).[4] In contrast, for the HPV issue, the logic-based correction was seen as more credible among those dismissive (p=.01) and the undecided (p=.10), but not significantly different from the humor-based correction among the convinced (p=.70). Thus, for HPV it appears that the logic-based correction is more palatable to those with low attitude accuracy, whereas the logic-based correction is more appealing to those

[4] For climate change, we only compare the convinced and undecided, as there was substantial skew in our measure. Only 17% of participants reported 50% or less agreement with the statement that human activity is causing climate change in the pre-test, so we cannot examine the dismissive separately.

with *high* attitude accuracy on the issue for climate change and gun control.

Third, we consider the effects on attitude change. For the climate change issue, the effects of corrections depend on pre-existing beliefs, with a significant higher order interaction, $F(2,128)=7.19$, $p=.001$, R^2 change=.095, as suggested by two significant interactions between misperceptions and exposure to the humor-based ($b=.26$, *S.E.*=.08, $p=.001$) and logic-based corrections ($b=.28$, *S.E.*=.08, $p<.001$) replies. The results suggest that among the undecided on climate change, exposure to the misinformation (absent correction) produced a shift towards *more* accurate attitudes, while exposure to either the logic-based or humor-based correction neutralize this effect (see Figure 4a). We see no effects among the convinced.

For the gun control issue, we continue to observe no overall effects on change in attitudes, even when accounting for pre-existing misperceptions (see Figure 4b).

Finally, for the HPV vaccination issue—which above showed the most promising effects of correction—we find that pre-existing misperceptions condition response to the logic-based correction ($b=-.31$, *S.E.*=.16, $p=.05$), leading to a marginally significant higher order interaction, $F(2, 132)=2.91$, $p=.06$, R^2 change = .033. The results align with our expectations in H1b—among the dismissive ($p<.001$) and undecided ($p=.01$) on the HPV vaccine, exposure to the logic-based correction produced more accurate attitudes compared to the misinformation condition. Among the convinced, however, the pattern of results is different (see Figure 4c), with the humor-based correction mitigating a drop in accurate beliefs that occurs among misinformation-only or the logic-based correction, although the pairwise differences do not reach statistical significance.

Discussion

This study examines whether two types of corrections—highlighting and refuting the logical fallacies in misinformation as suggested by inoculation research using either a logic-

based or humor-based approach—can function as a form of observational correction on social media, leading people to update their attitudes on controversial issues. We test this proposition across three issues from diverse domains: health, science, and politics. Our results suggest that corrections using this critical thinking-based approach have promise, but the effects depend on the issue domain, the type of inoculation–humor versus logic–strategy, and pre-existing misperceptions on the issue.

Overall, our corrections were most successful for misinformation on the HPV issue. Both the logic-based and the humor-based corrections were effective in leading individuals to report greater agreement with expert consensus that the HPV vaccine does not cause auto-immune disorders. The logic-based correction appeared more effective, boosting accuracy by 16 percentage points for the entire sample. Importantly, these effects appeared largely centered upon those who originally held stronger misperceptions on the issue, on average moving the dismissive into the "undecided" category when receiving the logic-based correction.

Our observations of credibility assessments may provide insight into the mechanism for these effects. The dismissive found the logic-based correction more credible than the humor-based correction – and seeing the logic-based correction depressed the credibility of the misinformation tweet for this group. Altogether, the logic-based correction was seen as most credible, reduced the credibility of the misinformation, and led to more accurate issue attitudes among the dismissive, highlighting the important role credibility likely played.

Results among the convinced, meanwhile, provide evidence for the merits of the humor-based correction. The humorous cartoon negated a drop in attitudinal accuracy as compared to the misinformation-only condition for the convinced. In contrast, the logic-based correction *boosted* credibility perceptions of the misinformation tweet among the convinced, suggesting a

backfire effect may have occurred for this group (e.g., Nyhan & Reifler, 2010; but also Wood & Porter, 2016). If so, this backfire effect did not change their minds completely; instead the average drop in attitude accuracy likely reflects greater uncertainty about the effects of the HPV.

Unlike among the dismissive, the humor-based correction was seen as equally credible compared with the logic-based correction among the convinced. Convinced individuals may have been willing to spend the effort to understand a humorous cartoon they agreed with, strengthening their resistance to the misinformation message. Future research should explore the mechanisms behind the effectiveness of a humor-based approach, paying special attention to its appeal among those who agree with its content, who may be willing to engage more deeply with the complex message. However, humor may be a less successful correction strategy for those predisposed to resist its message. Alternatively, humor-based messages may have other benefits – such as attention or virality—that are not studied here.

There are several reasons why the HPV vaccination misinformation may have been especially amenable to corrections employing critical thinking strategies. First, the misinformation tweet was seen as the *least* credible in the HPV condition, regardless of whether participants saw a correction. As such, the critical thinking strategy in the correction had an easier target to criticize, which may have boosted its persuasive power. Second, the corrective responses for the HPV issue were also seen as relatively more credible than the climate change issue (although equally credible as the gun control issue), which may have contributed to their ability to change attitudes (e.g., Hovland & Weiss, 1951). Third, participants rated the HPV

vaccine as less personally relevant than the other two issues.[5] Deeply entrenched and personally relevant issues are more likely to produce motivated reasoning, inhibiting our ability to correct misperceptions on these topics (Bode & Vraga, 2015; Taber & Lodge, 2006).

So why didn't the refutation strategies work for the gun control or climate change issues? For gun control, we see little evidence of change in issue attitudes or credibility perceptions, even when accounting for pre-existing beliefs on the issue. For this prominent and explicitly partisan issue, attitude change may be quite difficult.

For climate change, the results are even more problematic, as it appears that those undecided on climate change reported *more accurate* attitudes after seeing the misinformation-only tweet; this gain was neutralized when the misinformation was corrected using either a logic-based or humor-based strategy. The fact that a misinformation tweet on its own seemingly produced a backfire effect leading to *more accurate* attitudes is troubling and merits further research. This counterintuitive response may be a result from the skew in issue attitudes on climate change, which we discuss subsequently.

This study has a number of limitations. First, our sample was not representative of the U.S. population. While this likely does not undermine our effects – research suggests Mechanical Turk samples produce similar effects as population samples (e.g., Coppock, Leeper, & Mullinix, 2018; Wood & Porter, 2016) – it does limit our ability to examine the climate change topic in

[5] A pre-test question asked participants to rate the personal importance of eight issues on five-point scales from very unimportant to very important. The HPV vaccine was rated less important (M=2.92, $S.E.$=1.22) than either gun control (M=4.01, $S.E.$=1.12, t=14.78, p<.001) or climate change (M=3.94, $S.E.$=1.23, t=13.41, p<.001).

particular. We are only able to examine effects among the undecided and convinced in our sample; we remain uncertain about how critical thinking-based corrections might work among those who are skeptical of climate change. Second, our misinformation messages rarely reduced accurate beliefs for any of the issues studied. Inoculation-based strategies may work differently when the original misinformation is more powerful; a proposition that future research should test. Third, our corrections used critical thinking strategies that are largely embedded in the visual cues of the correction (e.g., the cartoon versus logic graphic), as well as in the text. Therefore, we cannot tell whether text or visual cues contributed to its effects (or lack thereof).

Fourth, our measures of message credibility are adapted from previous work on user corrections (Vraga & Bode, 2018), but are limited in scope and not directly comparable to other common measures of credibility (e.g., Flanagin & Metzger, 2007). We also cannot speak to people's perceptions of the credibility of the *source* of the misinformation and correction. Manipulations of source cues–such as gender or expertise–may intersect with message features and audience predispositions to condition effects. Fifth, our results are cross-sectional in nature, and we cannot speak to whether the effects of correction will endure over time. This is especially problematic for the humor-based approaches; it is possible that these humorous messages will produce sleeper effects and have stronger impact over time (Nabi, Moyer-Gusé, & Byrne, 2007). Future research should test the longitudinal effects of misinformation. Finally, although we attempt to test the effects of these messages across different issue domains, we cannot separate whether differences arise from the issue itself (e.g., health vs. politics) or the type of logical fallacy invoked (e.g., correlation implies causation vs. oversimplification).

Overall, this paper combines two promising approaches to addressing misinformation on social media: inoculation and observational correction. Both a logic-based and humor-based

corrections that use the strategy of inoculation campaigns to uncover flaws in reasoning in misinformation have merit in responding to misinformation about the HPV vaccine on social media, although these results do not translate into the more partisan issues of gun control and climate change. Those attempting to address online misinformation should carefully consider the nature and context of the misinformation and the target audience in creating effective interventions that can be shared among the general public.

References

Allcott, H., Gentzkow, M., & Yu, C. (2018). Trends in the diffusion of misinformation on social media. *arXiv preprint arXiv:1809.05901*.

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59-79. Doi: 10.1177/1077699015606057

Banas, J., & Rains, S. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281-311. Doi: 10.1080/03637751003758193

Banas, J., & Richards, A. S. (2017). Apprehension or motivation to defend attitudes? Exploring the underlying threat mechanism in inoculation-induced resistance to persuasion. *Communication Monographs, 84*(2), 164-178. Doi: 10.1080/03637751.2017.1307999

Becker, A. B. (2018). Live From New York, It's Trump on Twitter! The Effect of Engaging With Saturday Night Live on Perceptions of Authenticity and the Salience of Trait Ratings. *International Journal of Communication*, *12*, 22. Doi: 1932–8036/20180005

Becker, A. B., & Bode, L. (2018). Satire as a source for learning? The differential effect of news versus satire exposure on net neutrality knowledge gain. *Information, Communication, and Society, 21,* 612-625. Doi: 10.1080/1369118X.2017.1301517

Becker, A. B., & Waisman, D. J. (2013). From funny features to entertaining effects: Connecting approaches to communication research on political comedy. *Review of Communication,* 13(3), 161-183. Doi: 10.1080/15358593.2013.826816

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, *10*(2), e0118093. Doi: 10.1371/journal.pone.0118093

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. Doi: 10.1111/jcom.12166

Bolsen, T., & Druckman, J. N. (2015). Counteracting the politicization of science. *Journal of Communication*, *65*(5), 745-769. Doi: 10.1111/jcom.12171

Bore, I. L. K., & Reid, G. (2014). Laughing in the face of climate change? Satire as a device for engaging audiences in public debate. *Science Communication*, *36*(4), 454-478. Doi: 10.1177/1075547014534076

Brewer, P. R., & McKnight, J. (2015). Climate as comedy: The effects of satirical television news on climate change perceptions. *Science Communication*, *37*(5), 635-657. Doi: 10.1177/1075547015597911

Brewer, P. R., McKnight, J. (2017). "A statistically representative climate change debate": Satirical television news, scientific consensus, and public perceptions of global warming. *Atlantic Journal of Communication*, *25*, 166-180. Doi: 10.1080/15456870.2017.1324453

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., ... & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378-1384. Doi: 10.2105/AJPH.2018.304567

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., ... & Sandhu, M. (2019). Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior*, 1-23. Doi: 10.1007/s11109-019-09533-0

Compton, J. (2005). Comparison, contrast, and synthesis of Aristotelian rationality and

inoculation. *Journal of the Speech and Theatre Association of Missouri, 35*, 1-23.

Compton, J. (2017, November). Prophylactic versus therapeutic inoculation treatments for resistance to influence. *Social Cognition Division, National Communication Association*: Dallas, TX.

Compton, J., Jackson, B., & Dimmock, J. A. (2016). Persuading others to avoid persuasion: Inoculation theory and resistant health attitudes. *Frontiers in Psychology*, *7*, 122. Doi: 10.3389/fpsyg.2016.00122

Compton, J., & Ivanov, B. (2013). Vaccinating voters: Surveying political campaign inoculation scholarship. In *E. L. Cohen (Ed.), Communication yearbook 37* (pp. 250-283). New York, NY: Routledge.

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences,* online first. doi: 10.1073/pnas.1808083115

Cook, J., Ellerton, P., and Kinkead, D. (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters*, *11*(2).

Cook, J., Lewandowsky, S., & Ecker, U. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE*, *12*(5): e0175799. DOI: 10.1371/journal.pone.0175799

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldareli, G., Stanley, H. E., & Quattrociocchi, W. (2017). The spreading of misinformation online. *Proceedings of the National Academy of Sciences, 113,* 554-559. Doi: 10.1073/pnas.1517441113

Ecker, U. K. H., Lewandowsky, S., Cheung, C. S. C., & Maybery, M. T. (2015). He did it! She did it! No, she did not! Multiple causal explanations and the continued influence of

misinformation. *Journal of Memory and Language*, *85*,101–115. Doi: 10.1016/j.jml.2015.09.002

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*, 1087–1100. Doi: 10.3758/MC.38.8.1087

Farinella, M. (2018). The potential of comics in science communication. *Journal of Science Communication*, *17*(01), Y01-1. Doi: 10.22323/2.17010401

Flanagin, A. J., & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*, *9*, 319-342. Doi: 10.1177/1461444807075015

Garrett, R. K., & Weeks, B. E. (2013). The promise and peril of real-time corrections to political misperceptions. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 1047-1058). ACM. Doi: 10.1145/2441776.2441895

Guidry, J. P., Carlyle, K., Messner, M., & Jin, Y. (2015). On pins and needles: how vaccines are portrayed on Pinterest. *Vaccine*, *33*(39), 5051-5056. Doi: 10.1016/j.vaccine.2015.08.064

Harvey, J. A., Van den Berg, D., Ellers, J., Kampen, R., Crowther, T. W., Roessingh, P., ... & Stirling, I. (2017). Internet blogs, polar bears, and climate-change denial by proxy. *BioScience*, *68*(4), 281-287. Doi: 10.1093/biosci/bix133

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635-650.

Jee, B. D., & Anggoro, F. K. (2012). Comic cognition: exploring the potential cognitive impacts

of science comics. *Journal of Cognitive Education and Psychology*, *11*(2), 196-208. DOI: 10.1891/1945-8959.11.2.196

Juthe, A. (2009). Refutation by parallel argument. *Argumentation*, *23*(2), 133. Doi: 10.1007/s10503-008-9109-8

Kim, Y., & Chen, H. T. (2016). Social media and online political participation: The mediating role of exposure to cross-cutting and like-minded perspectives. *Telematics and Informatics*, *33*(2), 320-330. Doi: 10.1016/j.tele.2015.08.008

Kim, Y., Chen, H. T., & gil de Zúñiga, H. (2013). Stumbling upon news on the Internet: Effects of incidental news exposure and relative entertainment use on political engagement. *Computers in Human Behavior*, *29*(6), 2607-2614. Doi: 10.1016/j.chb.2013.06.005

Kim, Y. M., & Vishak, J. (2008). Just laugh! You don't need to remember. The effects of entertainment media on political information acquisition and information processing in political judgments. *Journal of Communication, 58,* 338-360. Doi: 10.1111/j.1460-2466.2008.00388.x

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the currency of democratic citizenship. *The Journal of Politics*, *62*(3), 790-816. Doi: 10.1111/0022-3816.00033

Levin, S. (2017, May 16). Facebook promised to tackle fake news. But the evidence shows it's not working. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2017/may/16/facebook-fake-news-tools-not-working

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and

coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369. Doi: 10.1016/j.jarmac.2017.07.008

Lewandowsky, S., Ecker, U. K., 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. Doi: 10.1177/1529100612451018

Madden, K., Nan, X., Briones, R., & Waks, L. (2012). Sorting through search results: a content analysis of HPV vaccine information online. *Vaccine*, *30*(25), 3741-3746. Doi: 10.1016/j.vaccine.2011.10.025

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on Twitter: when do corrections have an effect? *Political Communication*, *35*(2), 196-219. Doi: 10.1080/10584609.2017.1334018

McCright, A. M., Charters, M., Dentzman, K., & Dietz, T. (2016). Examining the Effectiveness of Climate Change Frames in the Face of a Climate Change Denial Counter-Frame. *Topics in Cognitive Science*, *8*(1), 76-97. Doi: 10.1111/tops.12171

McGuire, W. J., & Papageorgis, D. (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *Public Opinion Quarterly*, *26*, 24-34. Doi: 10.1037/h0042026

Mihailidis, P., & Viotty, S. (2017). Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American Behavioral Scientist*, *61*(4), 441-454. Doi: 10.1177/0002764217701217

Nabi, R. L., Moyer-Gusé, E., & Byrne, S. (2007). All joking aside: A serious investigation into the persuasive effect of funny social issue messages. *Communication Monographs*, *74*(1), 29-54. Doi: 10.1080/03637750701196896

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303-330. Doi: 10.1007/s11109-010-9112-2

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine*, *33*(3), 459-464. Doi: 10.1016/j.vaccine.2014.11.017

Pew Research Center. (2018). Social media fact sheet. Retrieved from http://www.pewinternet.org/fact-sheet/social-media/

Pew Research Center. (2017). The future of truth and misinformation online. Retrieved from: http://www.pewinternet.org/2017/10/19/the-future-of-truth-and-misinformation-online/

Pennycook, G., & Rand, D. G. (2017). Who falls for fake news? The roles of analytic thinking, motivated reasoning, political ideology, and bullshit receptivity. *SSRN Electronic Journal*.

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? *The Psychology of Learning and Motivation*, *41*, 265–292. Doi: 10.1016/S0079-7421(02)80009-3

Shao, C., Hui, P. M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PloS one*, *13*(4), e0196087. Doi: 10.1371/journal.pone.0196087

Skurka, C., Niederdeppe, J., Romero-Canyas, R., Acup, D. (2018). Pathways of influence in emotional appeals: Benefits and tradeoffs of using fear or humor to promote climate change-related intentions and risk perceptions. *Journal of Communication*, *68*, 169-193. Doi: 10.1093/joc/jqx008

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political

beliefs. *American Journal of Political Science*, *50*(3), 755-769. Doi: 10.1111/j.1540-5907.2006.00214.x

Tan, A. S., Lee, C. J., & Chae, J. (2015). Exposure to health (mis) information: Lagged effects on young adults' health behaviors and potential pathways. *Journal of Communication*, *65*(4), 674-698. Doi: 10.1111/jcom.12163

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*(3), 460-480. Doi: 10.1080/10584609.2015.1102187

Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. Available at SSRN: http://ssrn.com/abstract=3144139

Tully, M., Vraga, E. K., & Bode, L. (2019). Designing and Testing News Literacy Messages for Social Media. *Mass Communication and Society*, 1-25. Doi: 10.1080/15205436.2019.1604970

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2). Doi: 10.1002/gch2.201600008

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, *39*(5), 621-645. Doi: 10.1177/1075547017731776

Vraga, E. K., & Bode, L. (2018). I do not believe you: how providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, *21*(10), 1337-1353. Doi: 10.1080/1369118X.2017.1313883

Vraga, E. K., Johnson, C. N., Carr, D. J., Bode, L., & Bard, M. T. (2014). Filmed in front of a live studio audience: Laughter and aggression in political entertainment programming.

Journal of Broadcasting & Electronic Media, *58*(1), 131-150. Doi: 10.1080/08838151.2013.875020

Wood, T., & Porter, E. (2016). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 1-29. Doi: 10.1007/s11109-018-9443-y

Young, D. G. (2008). The privileged role of the late-night joke: Exploring humor's role in disrupting argument scrutiny. *Media Psychology*, *11*, 119-142. Doi: 10.1080/15213260701837073

Young, D. G., Jamieson, K. H., Poulsen, S., & Goldring, A. (2018). Fact-checking effectiveness as a function of format and tone: Evaluating FactCheck. org and FlackCheck. org. *Journalism & Mass Communication Quarterly*, *95*(1), 49-75. Doi: 10.1177/1077699017710453

Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S. & Quattrociocchi, W. (2015). Debunking in a world of tribes. arXiv preprint arXiv:1510.04267 doi: 10.1371/journal.pone.0181821

Figures

Figure 1: Interaction between correction and topic on change in issue attitude accuracy

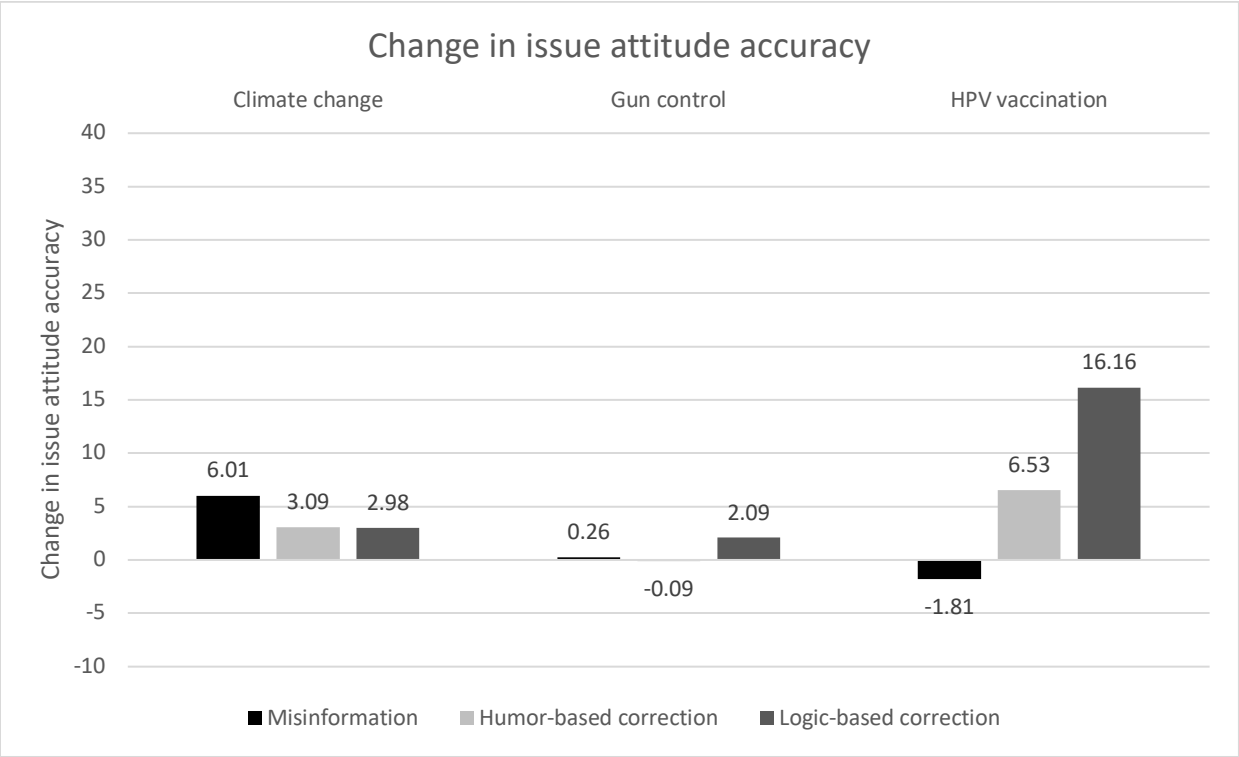


Figure 2: Interaction between correction and pre-existing beliefs on credibility of the original misinformation tweet for HPV vaccination

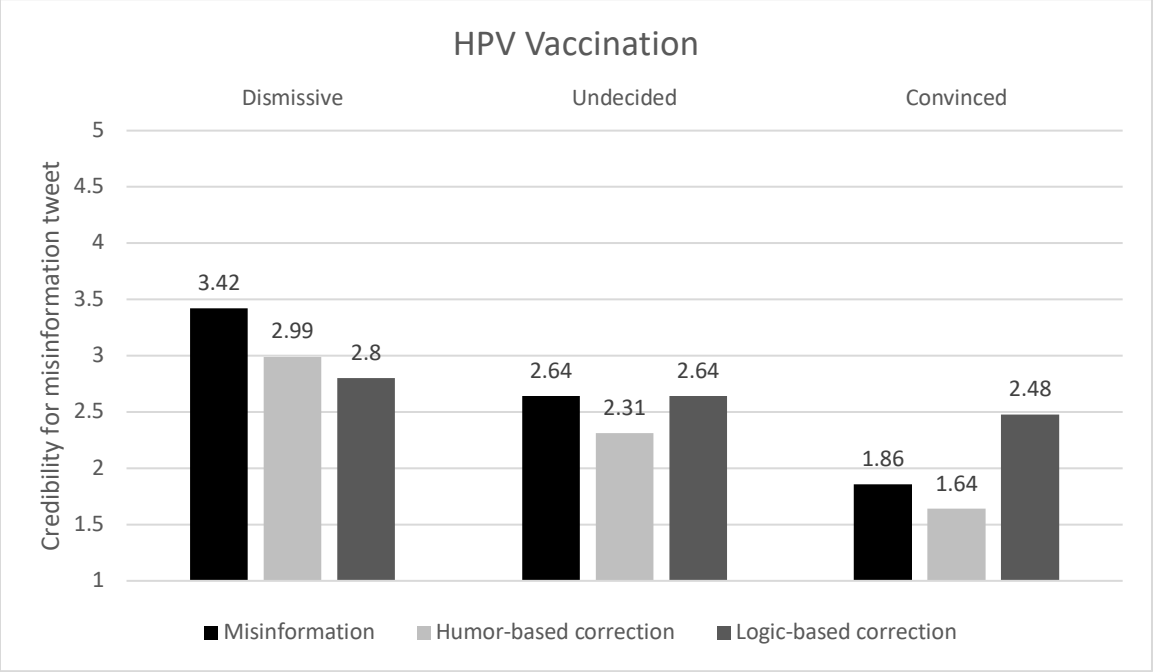
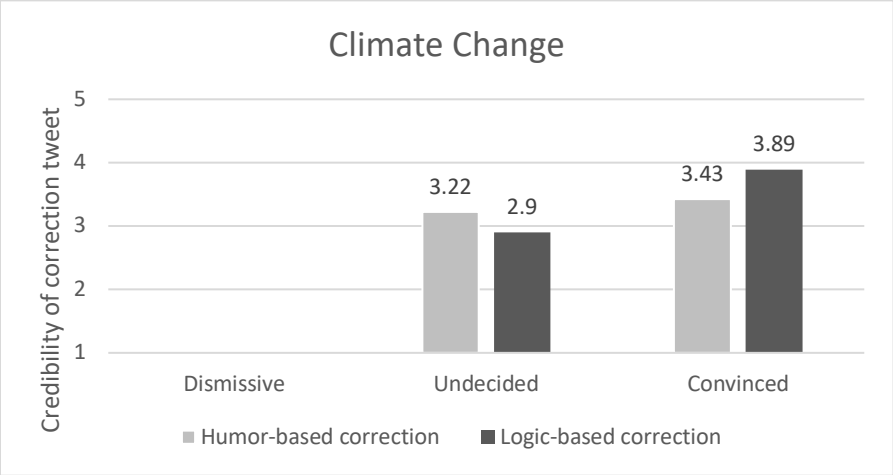
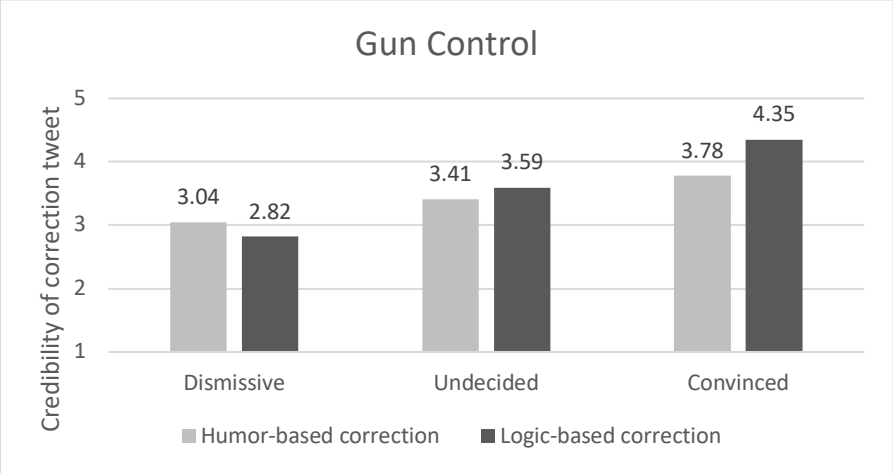


Figure 3: Interaction between correction and pre-existing beliefs on credibility of the correction tweet for (a) climate change, (b) gun control, and (c) HPV vaccination.





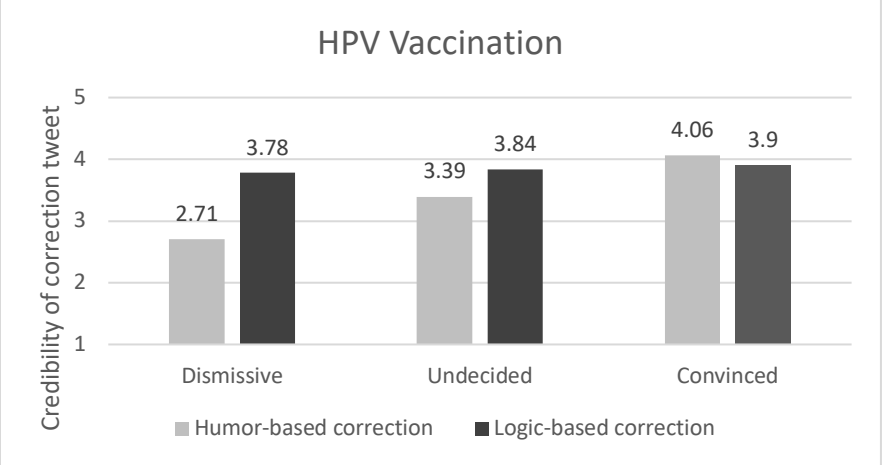
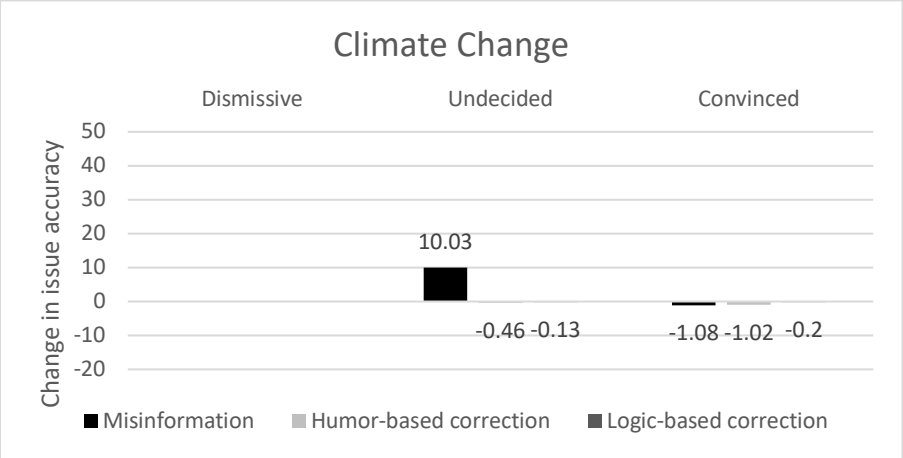
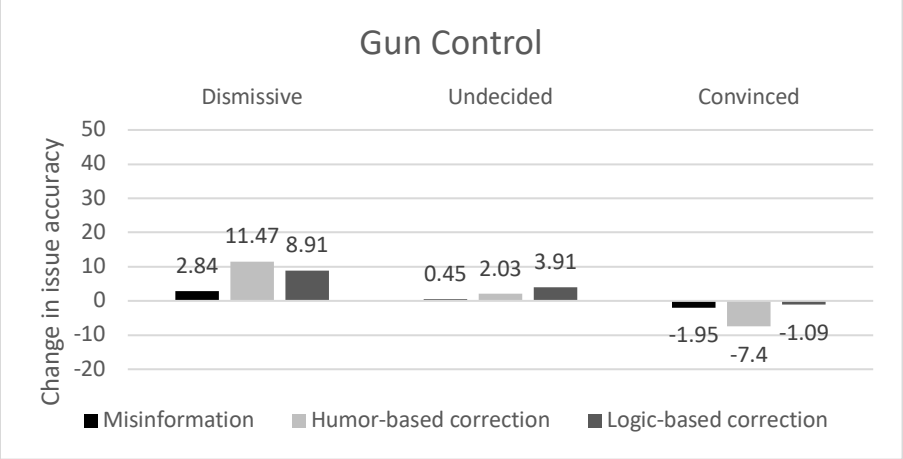
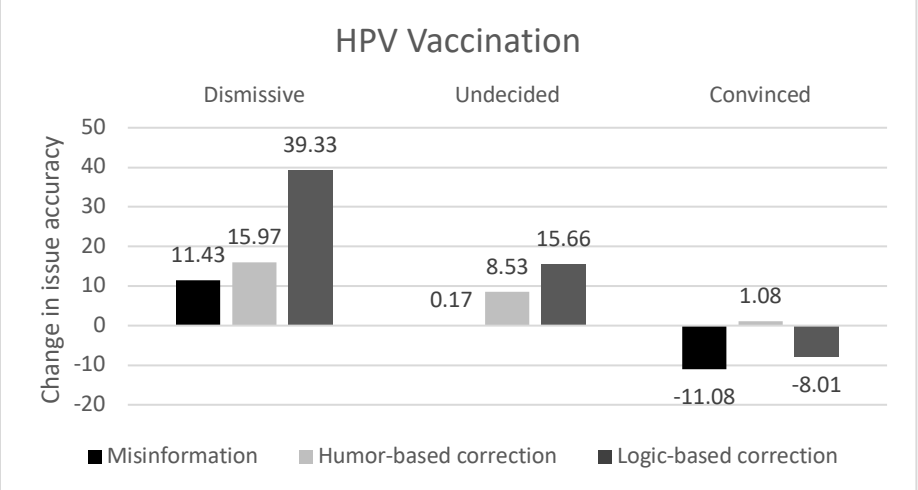


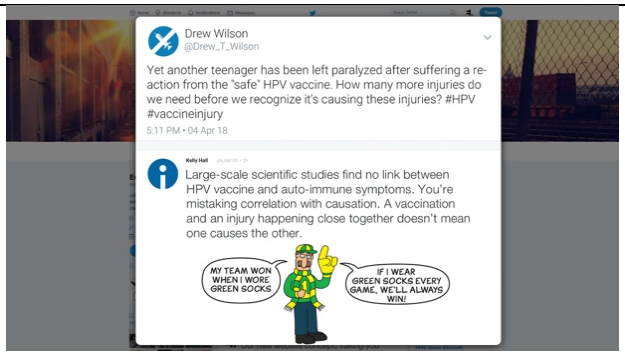
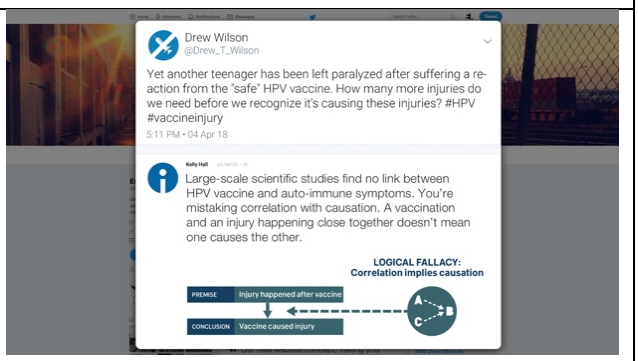
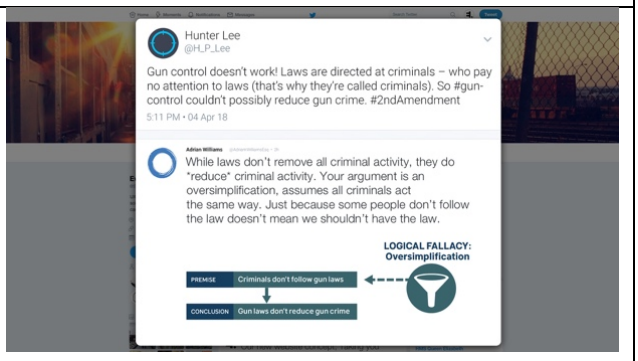
Figure 4: Interaction between correction and pre-existing beliefs on change in (a) climate change, (b) gun control, and (c) HPV vaccination attitudes.







Appendix A: Experimental stimuli

	Humor-based correction	Logic-based correction
Climate change		
HPV		
Gun control		

The misinformation-only condition includes the same original tweet absent the corrective response.