COGITATIO

Media and Communication (ISSN: 2183–2439)
2021, Volume 9, Issue 1, Pages 120–133
DOI: 10.17645/mac.v9i1.3519

Article

You're Definitely Wrong, Maybe: Correction Style Has Minimal Effect on Corrections of Misinformation Online

Cameron Martel [1,*], Mohsen Mosleh [1,2] and David G. Rand [1,3]

[1] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142, USA;
E-Mails: cmartel@mit.edu (C.M.), mmosleh@mit.edu (M.M.), drand@mit.edu (D.G.R.)
[2] Science, Innovation, Technology, and Entrepreneurship Department, Business School, University of Exeter, Exeter, EX4 4PU, UK
[3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

* Corresponding author

Submitted: 30 July 2020 | Accepted: 8 September 2020 | Published: 3 February 2021

Abstract
How can online communication most effectively respond to misinformation posted on social media? Recent studies examining the content of corrective messages provide mixed results—several studies suggest that politer, hedged messages may increase engagement with corrections, while others favor direct messaging which does not shed doubt on the credibility of the corrective message. Furthermore, common debunking strategies often include keeping the message simple and clear, while others recommend including a detailed explanation of why the initial misinformation is incorrect. To shed more light on how correction style affects correction efficacy, we manipulated both correction strength (direct, hedged) and explanatory depth (simple explanation, detailed explanation) in response to participants from Lucid ($N = 2{,}228$) who indicated they would share a false story in a survey experiment. We found minimal evidence suggesting that correction strength or depth affects correction engagement, both in terms of likelihood of replying, and accepting or resisting corrective information. However, we do find that analytic thinking and actively open-minded thinking are associated with greater acceptance of information in response to corrective messages, regardless of correction style. Our results help elucidate the efficacy of user-generated corrections of misinformation on social media.

Keywords
cognitive reflection test; corrections; dark participation; debunking; fake news; misinformation; social media

Issue
This article is part of the issue "Dark Participation in Online Communication: The World of the Wicked Web" edited by Thorsten Quandt (University of Münster, Germany).

1. Introduction

An estimated 3,6 billion people use social media as of 2020, with this number expected to only increase in the next decade (Clement, 2020). Furthermore, people are increasingly utilizing social media platforms as a primary source of news consumption—indeed, it has been estimated that about two-thirds of American adults at least occasionally get news via social media, despite apprehensions about the accuracy of such news (Shearer & Matsa, 2018). The advent of social media as a means of news dissemination has led to widespread concern over the spread of misinformation and 'fake news' ("fabricated information that mimics news media content in form but not in organizational process or intent"; Lazer et al., 2018, p. 1094). Although fake news comprises a relatively small proportion of Americans' daily media diet (0.15%; see Allen, Howland, Mobius, Rothschild, & Watts, 2020), it may still be harmful. For instance, in the months leading up to the 2016 USA Presidential election, false news stories favoring Trump were shared about 30 million times on Facebook; those favoring

Clinton were shared 8 million times (Allcott & Gentzkow, 2017). More recently, misinformation and disinformation about Covid-19 has spread quickly on social media (Frenkel, Alba, & Zhong, 2020), potentially with fatal consequences. As a result, there is great interest in identifying approaches to combat misinformation.

1.1. Combatting Misinformation at the Platform-Level

One approach is to implement platform-level interventions (i.e., efforts implemented by social media platforms that may be applied to all users). The most widely implemented such approach, applying fact-check tags on disputed or false-rated news items, has substantial limitations. Professional fact-checkers cannot possibly keep up with the pace at which misinformation is produced. In addition to limiting the reach of fact-checking, this may promote increased perceptions of accuracy for unlabeled false headlines ('implied truth effect'; Pennycook, Bear, Collins, & Rand, 2020). Relatedly, general warnings instructing users to be cautious about the accuracy of news content they read and share may result in decreased belief in true news stories ('tainted truth effect'; Clayton et al., 2019). Approaches based on inoculation (i.e., preemptive exposure to and warnings of fake news; Roozenbeek & van der Linden, 2019) and accuracy nudges (i.e., reminding people to think about accuracy before consuming or sharing news content; Pennycook et al., in press; Pennycook, McPhetres, Zhang, Lu, & Rand, 2020), which induce people to be more discerning prior to their contact with misinformation, show substantial promise. So does utilizing layperson judgments (e.g., by harnessing the wisdom of crowds through users or contractors hired to provide quality ratings) to supplement machine learning approaches to misinformation detection (Epstein, Pennycook, & Rand, 2020; Kim, Tabibian, Oh, Schölkopf, & Gomez-Rodriguez, 2018; Pennycook & Rand, 2019a). However, it seems unlikely that platforms will ever be entirely able to control the misinformation problem.

1.2. Combatting Misinformation at the User-Level

In addition to interventions that can be applied by the platforms, it is therefore important to determine what kind of user-generated corrections may be most effective at combatting misinformation online. While correcting misinformation may ideally be a source of positive participation online, it may easily devolve into unproductive and even harmful discourse. This gives rise to the question of what type of corrective message most effectively combats dark participation, rather than gives way to it? One dimension by which corrective messages may differ is that of strength—how forcefully the corrective message corrects the shared misinformation. Less forceful, hedged messaging may lead to increased engagement with and acceptance of the corrective message. For instance, Lewandowsky, Ecker, Seifert, Schwarz, and Cook (2012) argue that effective corrections should attempt to affirm the worldview and identity of the individual being corrected—thus, a hedged correction may be less abrasive towards the corrected individual and their worldview and identity. Furthermore, Tan, Niculae, Danescu-Niculescu-Mizil, and Lee (2016) analyzed effective corrective discourse on Reddit, and found that hedging (i.e., a message that indicated uncertainty, such as "it could be the case"; Tan et al., 2016, p. 622) was more common in more persuasive arguments. This is perhaps because hedging makes an argument easier to accept through the use of a softer tone (Lakoff, 1975). However, it has also been suggested that hedging may add increased uncertainty to the corrective message, thus reducing its efficacy (see Rashkin, Choi, Jang, Volkova, & Choi, 2017). This would suggest that more direct, less hedged corrections of misinformation may provide the clearest form of correction. Alternatively, recent evidence suggests that the tone of a correction (uncivil, affirmational, or neutral) may not affect the effectiveness of corrections to misinformation (Bode, Vraga, & Tully, 2020). Ultimately, there remains limited causal evidence as to whether and how correction strength may impact correction engagement.

Another dimension by which corrections may vary is by explanation depth; for instance, whether the debunking message consists of a simple negation, or includes an alternative account to fill in the gap left by correcting misinformation (see Lewandowsky et al., 2012). In favor of brief refutations, it has been argued that simple rebuttals of misinformation are most effective (Lewandowsky et al., 2012). However, others have argued to avoid simple negations (Nyhan & Reifler, 2012) and instead provide detailed debunking messages (see Chan, Jones, Hall Jamieson, & Albarracín, 2017). Thus, it remains unclear whether corrections should be simple negations of truth, or if they should contain a more detailed explanation of why the shared misinformation is false.

1.3. Current Research

In the current study, we investigate the causal role of different corrective messaging strategies on engagement with corrections. Using a survey experiment in which participants are presented with a series of social media posts, we induce most participants to indicate that they would share a false headline. We then manipulate the style of corrective message participants receive in response to their shared article. Corrective messages varied by strength (direct correction, hedged correction) and depth (simple explanation, detailed explanation). All corrections also included a (non-functional) link to a purported debunking article on *Snopes*, which should also increase the efficacy of the corrective message (see Vraga & Bode, 2018; for related research on 'snoping,' see Margolin, Hannak, & Weber, 2018).

We first predict that (H1) hedged corrections will be perceived as less aggressive and more polite than direct corrections, and that (H2) detailed corrections will be perceived as more informative and less unhelpful than simple corrections. We also predict that (H3) hedged, detailed corrections will elicit greater reply likelihood and (H4) predict greater acceptance of information, whereas direct and simple corrections will predict increased resistance of information. Finally, we anticipate that (H5) more analytic or actively open-minded individuals will have greater reply likelihood and acceptance of information in response to more detailed corrections.

The current research extends existing literature regarding debunking and corrections of fake news in three main ways. First, the existing literature assessing the effect of correction strength on correction engagement is primarily observational rather than causal (e.g., Tan et al., 2016). We seek to causally determine whether correction strength affects correction efficacy. Second, there is limited work assessing the interaction between various correction wording strategies. We assess not only whether there are main effects of correction strength and depth on engagement, but also if these correction styles may interact with one another. Third, we seek to explore the interaction between correction style and several key cognitive mechanisms which may impact the efficacy of certain forms of corrections. In particular, we utilize the cognitive reflection test (CRT; Frederick, 2005) to assess whether more analytic thinkers engage more with more detailed explanations. We also explore the role of actively open-minded thinking (Stanovich & West, 2007) in receptivity to various corrective messaging styles.

2. Methods

Our study was pre-registered at https://osf.io/eupwn/?view_only=cc6cd2cd0bae42788fcd28aacb505d9a. Furthermore, our full materials, data, and analysis code is available on the Open Science Framework (see https://osf.io/fvwd2/?view_only=cc6cd2cd0bae42788fcd28aacb505d9a).

2.1. Materials and Procedure

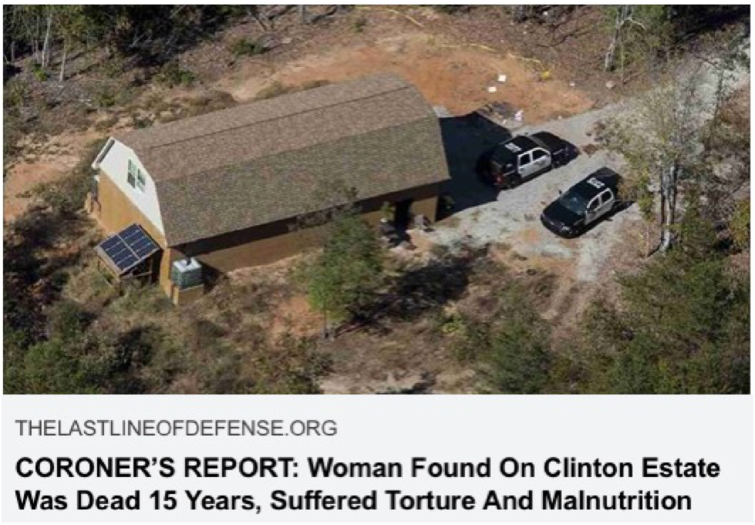
2.1.1. Participants

We recruited *N* = 2,228 participants (1,065 female, M_{age} = 44.84) via the online convenience sampling platform Lucid (Coppock & McClellan, 2019). Participants were first instructed to imagine they were currently on a social media platform such as Twitter or Facebook. Participants were then told they would be presented with a series of actual recent news headlines, as if they were appearing in their social media newsfeed.

2.1.2. News Headlines

Participants were randomly shown up to 28 actual headlines that appeared on social media, half of which were factually accurate (real news) and half of which were entirely untrue (fake news). Additionally, half of the headlines were favorable to the Democratic Party, and half were favorable to the Republican Party, based on pretest ratings (see Pennycook & Rand, 2019b). All fake news headlines were taken from *Snopes*. Real news headlines were selected from mainstream news sources (e.g., *NPR*, *The Washington Post*). Headlines were presented in the format of a social media post—namely, with a picture, headline, byline, and source (Figure 1).

After each headline, participants were asked whether or not they would share that article on social media publicly, such that other users could see and comment on it. If participants decided to share a real news article or decided not to share a fake news article, they were shown another headline. However, if participants decided to share a fake news article, then they proceeded to the rest of the study and saw no further headlines. Participants who did not share any fake news articles were not eligible to complete the correction message section of the study. This indication to share should simulate participants sharing such news articles as if they



THELASTLINEOFDEFENSE.ORG

CORONER'S REPORT: Woman Found On Clinton Estate Was Dead 15 Years, Suffered Torture And Malnutrition

Figure 1. Example news headline with picture, headline, byline, and source.

were actually on social media—indeed, recent research has found that self-reported willingness to share political news articles in online survey studies correlates with actual sharing on Twitter (Mosleh, Pennycook, & Rand, 2020).

2.1.3. Corrective Messages

Overall, 1,589 participants (71% of all participants) shared at least one fake news article, and thus completed the remainder of the study. After sharing a fake news article, participants were instructed to imagine receiving a public comment on their post. Participants were presented with one of four corrective messages, which varied by strength (direct, hedged) and depth (simple explanation, detailed explanation). These corrections were stylized as tweets from a fictional user. The first sentence of the message varied by strength—in the direct condition, the message read: "No way do I believe this article—it's definitely not true." In the hedged condition, the message read: "I'm not sure about this article—it might not be true." The second sentence of the message varied by depth—in the simple condition, the sentence read: "I found a link on Snopes that says this headline is false." In the detailed condition, the message read: "I found a link on *Snopes* that says this headline was created by a website that purposefully makes false stories." All messages ended with a stylized *Snopes* link (Figure 2).

2.1.4. Reply to Corrective Message

Next, participants were asked: "Would you reply to the above message?" 1 = "Yes, I would write a reply," 0 = "No, I would not write a reply." If participants indicated "Yes," they were asked to enter their reply via free response. If participants indicated "No," they were asked: "If you DID reply, what would you write?" and then allowed to enter their reply via free response.

2.1.5. Correction Motive

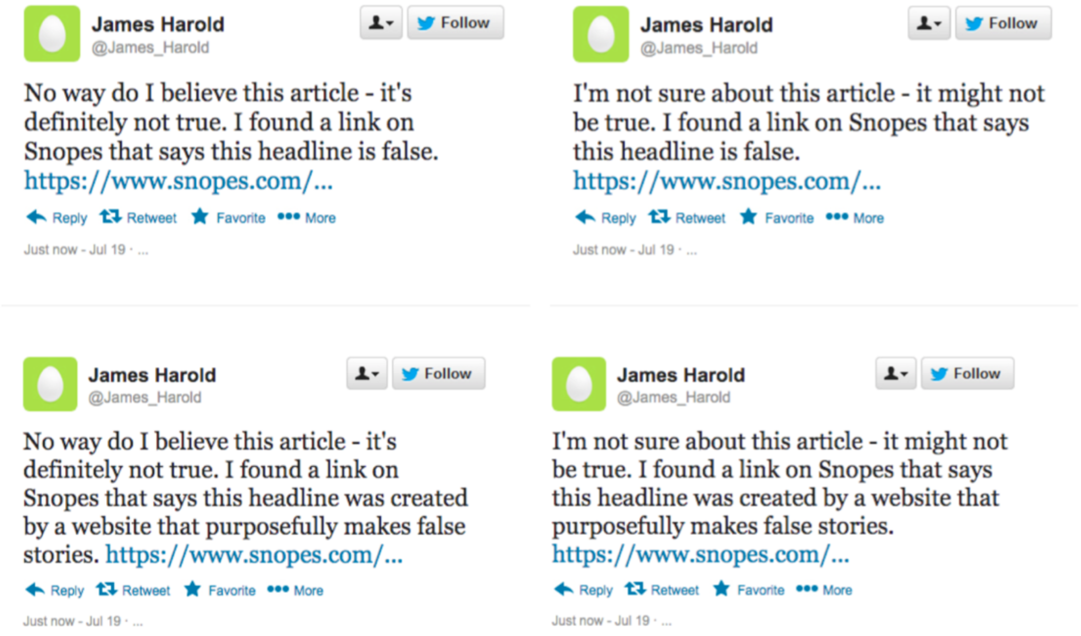
Participants were asked: "Why do you think the person wrote the message you received? Select all that apply." Participants could select from the following: "To inform me of valuable information," "To reinforce the image of themselves they'd like to present to me," "To develop a connection with me," "To achieve self-fulfillment for themselves," or "To get the word out about a specific cause.''

2.1.6. Evaluation of Corrector

After that, participants were asked to evaluate the trustworthiness of the person who wrote them the corrective message (Likert-scale: 1–7), as well as how positive and how negative their opinion was of the person who wrote them the message (Likert-scale: 1–7). Participants were also asked how much they agreed with the following statements: "The message I received on my social media post was [unhelpful/aggressive/informative/polite]." Likert-scale: 1 = Not at all agree, 7 = Strongly agree.

2.1.7. Self-Reported Belief-Updating

Then, participants were asked: "After viewing the comment on your shared article and replying to that comment, how do you view the accuracy of the article you



James Harold
@James_Harold

No way do I believe this article - it's definitely not true. I found a link on Snopes that says this headline is false. https://www.snopes.com/...

Just now - Jul 19 · ...

James Harold
@James_Harold

I'm not sure about this article - it might not be true. I found a link on Snopes that says this headline is false. https://www.snopes.com/...

Just now - Jul 19 · ...

James Harold
@James_Harold

No way do I believe this article - it's definitely not true. I found a link on Snopes that says this headline was created by a website that purposefully makes false stories. https://www.snopes.com/...

Just now - Jul 19 · ...

James Harold
@James_Harold

I'm not sure about this article - it might not be true. I found a link on Snopes that says this headline was created by a website that purposefully makes false stories. https://www.snopes.com/...

Just now - Jul 19 · ...

Figure 2. Corrective message conditions. Note: Clockwise, from top left: direct, simple; hedged, simple; hedged, detailed; direct, detailed.

shared?” Likert-scale: 1 = Much less accurate than initially thought, 2 = Slightly less accurate than initially thought, 3 = As accurate as initially thought, 4 = Slightly more accurate than initially thought, 5 = Much more accurate than initially thought.

2.1.8. Cognitive Reflection Test

The CRT is a brief task which measures participant tendencies to engage in analytic thinking. CRT items include an intuitive yet incorrect answer, which participants must override in order to answer correctly (e.g., “The ages of Mark and Adam add up to 28 years old. Mark is 20 years older than Adam. How many years old is Adam?” The common, intuitive answer is eight, whereas the correct answer upon reflection is four).

Participants completed a reworded version of the original CRT (Frederick, 2005; Shenhav, Rand, & Greene, 2012) and a four-item non-numeric CRT (Thomson & Oppenheimer, 2016).

2.1.9. Actively Open-Minded Thinking

Participants also completed a shortened version of the actively open-minded thinking scale (AOT; Stanovich & West, 2007). The AOT measures actively open-minded thinking, or the tendency to be open towards opinions or positions different from one’s own (e.g., “A person should always consider new possibilities.” 1 = I strongly disagree, 7 = I strongly agree).

2.1.10. Additional Measures

Participants next completed a brief political knowledge measure and standard demographics. Participants also were asked which social networking sites they use (Facebook, Twitter, Instagram, LinkedIn, Snapchat, Other). Participants were asked how often they are on social media and, if they indicated that they had either a Facebook or a Twitter account, how often they share content on Facebook or Twitter (“Never,” “Less often,” “Every few weeks,” “1 to 2 days a week,” “3 to 6 days a week,” “About once a day,” “2 to 5 times a day,” “5 to 10 times a day,” or “Over 10 times a day”).

2.2. Analysis of Free Response Replies

Following our main study, we recruited 819 participants from Amazon Mechanical Turk to crowdsource coding of the free response replies we collected on several dimensions. Free response replies were rated an average of 5.14 times each (*SD* = 2.24), and each participant rated 10 replies on seven key dimensions. Participants were first given instructions carefully detailing each rating category. Participants then evaluated each headline on these dimensions, via a Likert-scale (1–7). These seven dimensions of evaluating replies were informed by categories of responding to corrective information as detailed by Prasad et al. (2009). Detailed explanations of these seven dimensions (denying original belief, belief updating, counter-arguing, attitude bolstering, selective exposure, disputing rationality, and inferred justification) may be found in our Supplemental Materials here: https://osf.io/fvwd2/?view_only=cc6cd2cd0bae42788fcd28aacb505d9a.

2.2.1. Rating Procedure

Participants first read instructions detailing how they will be asked to evaluate replies to corrective messages using seven different categories of response types. Participants then read the descriptions of these response types, and answered a reading comprehension check. Participants who failed this comprehension check were presented with the response type descriptions a second time. Finally, participants viewed 10 different replies and rated each response by all seven response type categories. Participants also were asked: “Overall, how positive is this reply?” and “Overall, how negative is this reply?” Likert-scale: 1 = Not at all [positive/negative], 7 = Very [positive/negative]. Participants also evaluated whether the replier indicated they only shared the fake article as a joke, or if the replier indicated that they plan on looking up more information about the article they shared.

All written replies from the Lucid study were rated by Amazon Mechanical Turk raters, except replies which were either blank or simply said “nothing.” These replies were automatically coded as a 1 (not at all) across all categories. There were 320 such replies in total.

2.2.2. Intraclass Correlations

In order to assess the consistency of measurements made by our Amazon Mechanical Turk raters assessing the same replies, we computed intraclass correlations (ICC; descriptive measure of how strongly units in a group resemble one another) for each of the seven response type categories, plus ratings of overall positivity and negativity. In particular, we utilized a one-way random effects ICC model (since each reply was measured by a different set of randomly selected raters), as well as average measures, as our analyses ultimately utilize the average ratings for each reply (see Treventhan, 2017). Across all nine categories, our ICC1k was fair on average, meaning that reply ratings within response type categories adequately resembled one another, $ICC_{avg} = 0.46$ (common guidelines interpret greater than 0.40 as fair; Cicchetti & Sparrow, 1981), $ICC_{DenyOriginalBelief} = 0.33$, $ICC_{BeliefUpdating} = 0.57$, $ICC_{Counter\text{-}arguing} = 0.48$, $ICC_{AttitudeBolstering} = 0.36$, $ICC_{SelectiveExposure} = 0.38$, $ICC_{DisputingRationality} = 0.30$, $ICC_{InferredJustification} = 0.24$, $ICC_{Positive} = 0.73$, $ICC_{Negative} = 0.72$ (all *p*s < .001).

3. Results

3.1. Hedged Corrections Perceived as Less Aggressive, More Polite

In order to assess the effect of our correction style conditions on perceptions of corrections, we performed several analyses. We performed a linear regression model predicting how aggressive participants perceived the corrective messages they received, entering correction strength, depth, and their interaction as predictors. As expected, we found that participants who received hedged corrections perceived the correction as less aggressive, $b = -0.30$, $SE = 0.05$, $t(1558) = -6.19$, $p < .001$. There was no main effect of correction depth, nor interaction between conditions, $ps > .273$. Similarly, we found, as expected, that participants who received hedged corrections perceived the corrections as more polite, $b = 0.19$, $SE = 0.05$, $t(1557) = 4.00$, $p < .001$. Again, there was no main effect of correction depth nor interaction between conditions, $ps > .523$. Together, these measures suggest that there was a noticeable difference between direct and hedged corrective messages, such that hedged corrections were perceived as less aggressive and more polite, which supports our first hypothesis (H1). Indeed, these results suggest that our hedged condition was both definitionally manipulating hedging (i.e., via indicating uncertainty in wording by stating "I'm not sure"), as well as manipulating perceived aggressiveness and politeness of the correction. Additional analyses also suggest that hedged corrections promote slightly more positive perceptions of the corrector (see the Supplementary File).

We also performed a general linear model predicting how informative participants perceived the corrective message they received. Surprisingly, we found no main effects or interactions between correction conditions, $ps > .196$. We next performed a similar analysis substituting informativeness with unhelpfulness, but again found no main effects or interactions, $ps > .103$. Therefore, our results do not support our second hypothesis (H2), as we did not observe that participants evaluated corrections with more detailed explanatory depth as more informative or less unhelpful. These results suggest that while explanatory depth was definitionally manipulated in our design (i.e., the 'detailed explanation' correction contained information beyond a simple negation), it is not the case that explanatory depth manipulated the extent to which participants perceived the correction as informative or helpful.

3.2. No Meaningful Effect of Correction Strength and Depth on Reply Likelihood

The fraction of participants who said they would reply is shown in Figure 3.

For our main analysis, we then entered correction strength and depth, as centered dummies, plus their

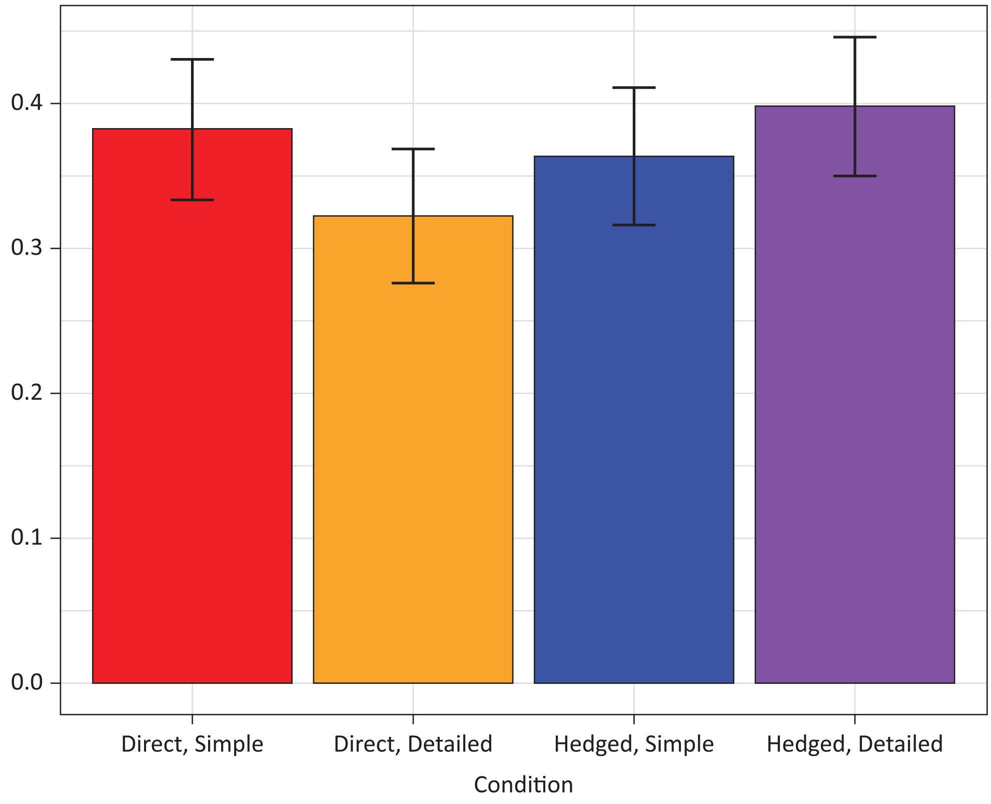


Figure 3. Likelihood of reply to the corrective message by condition. Notes: $N = 1{,}589$. Error bars reflect 95% confidence intervals.

interaction into a logistic regression to predict whether participants indicated that they would reply to the corrective message. Although we predicted that hedged and detailed corrections would elicit greater likelihood of replying, we found no main effect on reply likelihood of correction strength, $b = 0.06$, $SE = 0.05$, $z(1588) = 1.23$, $p = .219$, nor depth, $b = -0.03$, $SE = 0.05$, $z(1588) = -0.53$, $p = .598$ (Table 1).

Thus, our results do not support our third hypothesis (H3), that hedged, detailed corrections would elicit greater reply likelihood. We did find a (barely) significant interaction between correction strength and depth, $b = 0.10$, $SE = 0.05$, $z(1588) = 1.97$, $p = .049$, such that when the correction depth was detailed, hedged corrections elicited more responses than direct corrections. However, given our large sample and the fact that the p-value was only barely significant, this interaction should be interpreted with substantial caution.

3.3. No Meaningful Effect of Correction Strength and Depth on Reply Sentiment

As pre-registered, we averaged denying original belief and belief updating ratings to create a composite correction acceptance score (Figure 4).

We then entered correction strength and depth into a general linear model predicting correction acceptance score, allowing for an interaction between conditions. We found no significant main effects of correction strength, $b = 0.06$, $SE = 0.03$, $t(1588) = 1.96$, $p = .051$, or correction depth ($p = .600$) and no interaction between conditions ($p = .424$; Table 2). Our results did not support our fourth hypothesis (H4), which predicted that hedged, detailed corrections would elicit greater acceptance of information.

We next averaged the remaining five response categories (counter-arguing, attitude bolstering, etc.) into

Table 1. Logistic regression predicting likelihood of reply to corrective message.

	Estimate	Standard Error	*z*	*p*
Intercept	−0.55	0.05	−10.50	< .001***
Hedged	0.06	0.05	1.23	.219
Detailed Explanation	−0.03	0.05	−0.53	.598
Hedged*Detailed	0.10	0.05	1.97	.049*

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,588.

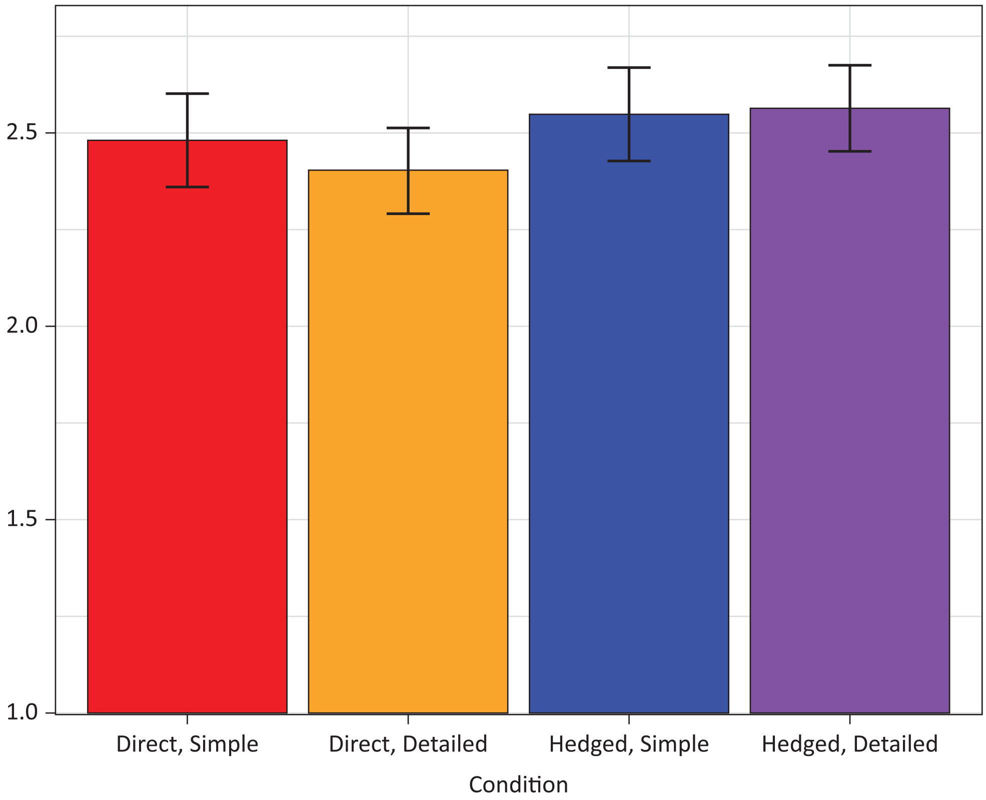


Figure 4. Average aggregated acceptance of corrective information (1–7 Likert-scale) by condition. Notes: $N = 1{,}589$. Error bars reflect 95% confidence intervals.

Table 2. General linear model predicting information acceptance by correction style condition.

	Estimate	Standard Error	*t*	*p*
Intercept	2.50	0.03	84.76	< .001***
Hedged	0.06	0.03	1.96	.051
Detailed Explanation	−0.02	0.03	−0.52	.600
Hedged*Detailed	0.02	0.03	0.80	.424

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,588.

an aggregated resisting information score (Figure 5), and predicted information resistance using a general linear model with correction strength and depth, allowing for an interaction.

We found no main effect of strength or depth, and no interaction between conditions, *ps* > .408 (Table 3).

We also performed five separate general linear models for each of the individual resisting information reply type categories. There were no significant main effects or interactions across all five linear models, *ps* > .146. Thus, our results again did not support our fourth hypothesis (H4), as direct and simple corrections did not predict increased resistance of information.

We also predicted overall positivity and overall negativity of reply using similar general linear models. Again, we found no significant main effects nor interactions, *ps* > .139.

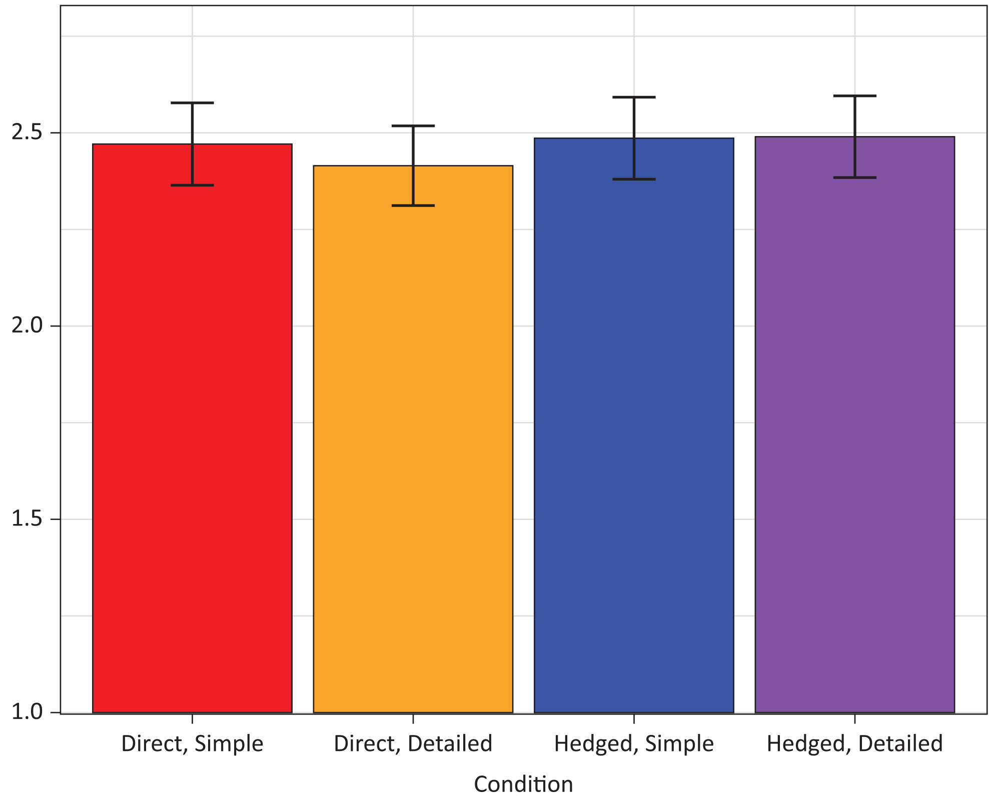


Figure 5. Average aggregated resistance of corrective information (1–7 Likert-scale) by condition. Notes: $N = 1,589$. Error bars reflect 95% confidence intervals.

Table 3. General linear model predicting information resistance by correction style condition.

	Estimate	Standard Error	*t*	*p*
Intercept	2.47	0.03	92.05	< .001***
Hedged	0.02	0.03	0.83	.408
Detailed Explanation	−0.01	0.03	−0.48	.632
Hedged*Detailed	0.01	0.03	0.56	.577

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,588.

3.4. No Effect of Correction Style on Self-Reported Belief Updating

We next predicted self-reported belief change (reverse-coded; 5 = Much less accurate than initially thought, 1 = Much more accurate than initially thought) using a general linear model, with correction strength and depth as predictors, allowing for their interaction (Figure 6). We found no significant main effect of correction strength, $b = 0.02$, $SE = 0.03$, $t(1551) = 0.75$, $p = .456$, or depth, $b = 0.03$, $SE = 0.03$, $t(1551) = 1.18$, $p = .240$, and no significant interaction between the conditions, $b = 0.05$, $SE = 0.03$, $t(1551) = 1.77$, $p = .078$.

3.5. Cognitive Reflection Predicts Increased Acceptance of Corrective Information

Next, we added CRT score as a predictor in our logistic regression predicting binary reply from correction strength and depth, allowing for all interactions. We found no significant main effect of, or interactions with, CRT score on reply likelihood, *ps* > .132. We then performed a similar analysis using a general linear model to predict aggregated acceptance of information. In this model, we found a notable main effect of CRT score, such that higher CRT score was associated with increased acceptance of corrective information, $b = 0.17$, $SE = 0.03$, $t(1587) = 5.77$, $p < .001$. We did not observe any significant interactions between CRT score and our correction conditions, contrary to our fifth hypothesis (H5) which predicted that more analytic participants would be more likely to accept detailed corrections (Table 4).

We next performed the same analysis except substituting accepting information with our composite resisting information score. Interestingly, we again found a positive main effect of CRT score, such that higher CRT score was associated with increased resistance of corrective information, $b = 0.06$, $SE = 0.03$, $t(1587) = 2.34$, $p = .019$; and no significant interactions between CRT score and our correction conditions (Table 5).

In order to further examine the relationship between CRT score and reply sentiment, we performed a *z*-test to compare the coefficient of CRT score on accepting information to the coefficient of CRT score on resisting information. We found that the coefficient of CRT score on accepting information was significantly greater than that of CRT score on resisting information, $z = 2.67$, $p = .008$. Our results thus suggest that on balance, participants with higher CRT scores are more accepting of corrective information.

3.6. Actively Open-Minded Thinking Predicts Increased Acceptance of Corrective Information

We again performed our main logistic regression model predicting binary reply, this time adding AOT score

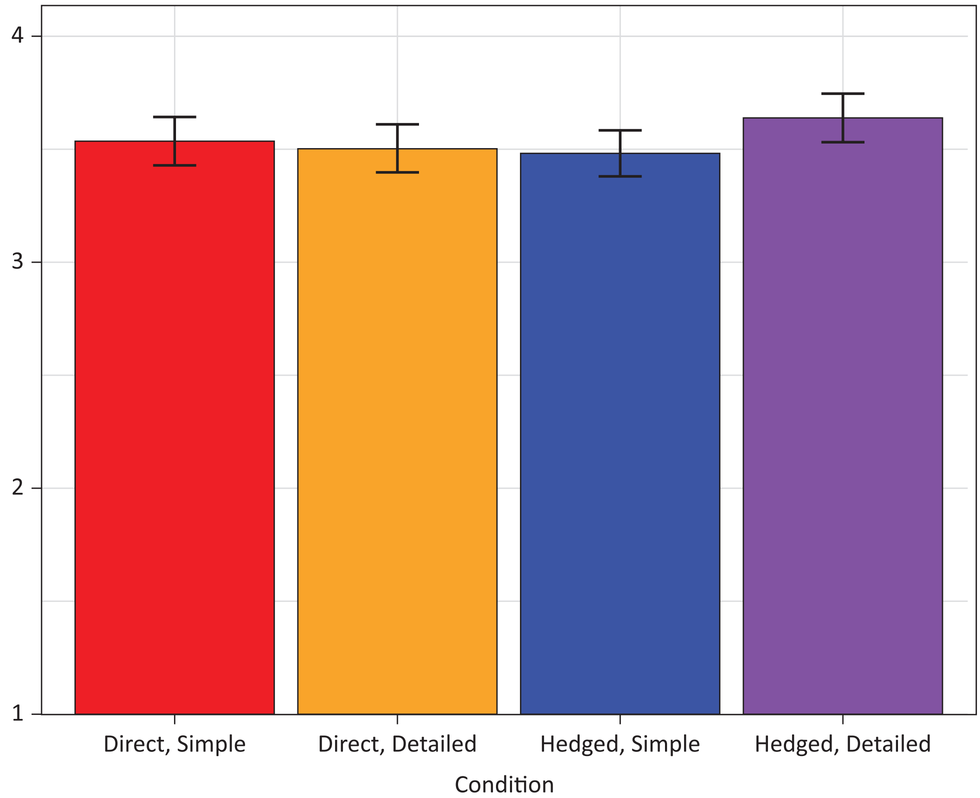


Figure 6. Self-reported belief updating (1–5 Likert-scale) by condition. Notes: $N = 1{,}552$. Error bars reflect 95% confidence intervals.

Table 4. General linear model predicting acceptance of information from correction style conditions and CRT as potential moderator.

	Estimate	Standard Error	*t*	*p*
Intercept	2.50	0.03	85.58	< .001***
Hedged	0.06	0.03	1.98	.048*
Detailed Explanation	−0.02	0.03	−0.59	.558
CRT Score	0.17	0.03	5.77	< .001***
Hedged*Detailed	0.02	0.03	0.74	.461
Hedged*CRT	0.04	0.03	1.53	.126
Detailed*CRT	−0.04	0.03	−1.20	.231
Hedged*Detailed*CRT	−0.02	0.03	−0.77	.443

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,587.

as a potential moderator, allowing for all interactions. We found a main effect of AOT on likelihood of reply, such that greater AOT score was associated with greater reply likelihood, $b = 0.15$, $SE = 0.05$, $z(1501) = 2.83$, $p = .005$. We found no interactions between correction conditions and AOT, *p*s > .494. We also ran a general linear model predicting aggregated acceptance of information from correction strength, depth, and AOT score, allowing for all interactions. We found a main effect of AOT score, such that greater AOT score was predictive of increased acceptance of corrective information, $b = 0.21$, $SE = 0.03$, $t(1501) = 7.00$, $p < .001$; and no significant interactions between correction conditions and AOT, *p*s > .353 (Table 6). These latter results thus do not support our fifth hypothesis (H5), as more actively open-minded participants were not more likely to accept information from detailed corrections.

We then performed the same analysis, substituting acceptance of information with aggregated resistance of information, but found no significant main effect of AOT ($p = .201$) and no significant interactions between correction conditions and AOT (*p*s > .243). Together, these results demonstrate that AOT is associated with increased acceptance, but not resistance, of corrective information.

We also performed several analyses looking at partisanship and social media use as potential moderators (see the Supplementary File).

4. Discussion

Our results suggest several conclusions about the effects of different styles of corrective messages on engagement with and replies to corrections of misinformation on social media. We find that hedged corrections are perceived as politer and less aggressive than direct corrections, and that hedged corrections result in a more positive perception of the corrector. Despite this, however, we do not find that hedged corrections are any more effective at eliciting replies to corrective messages, or promoting acceptance of corrective information. We consistently found no main effect of correction strength (direct, hedged) or explanatory depth (simple explanation, detailed explanation) on reply likelihood or reply sentiment. We did find some weak evidence of an interaction between correction strength and depth. This interaction was such that hedged, detailed corrections and direct, simple corrections yielded greater reply likelihood than direct, detailed corrections and hedged, simple corrections. This suggests that participants were perhaps sensitive to both correction strength and explanatory depth, yet neither correction style significantly impacted reply likelihood or the acceptance or rejection of the correction.

Overall, given our consistently minimal effects of correction strength and depth on responses to corrections, our findings suggest that correction style and wording

Table 5. General linear model predicting resistance of information from correction style conditions and CRT as potential moderator.

	Estimate	Standard Error	*t*	*p*
Intercept	2.47	0.03	92.07	< .001***
Hedged	0.02	0.03	0.82	.414
Detailed Explanation	−0.01	0.03	−0.52	.604
CRT Score	0.06	0.03	2.34	.019*
Hedged*Detailed	0.01	0.03	0.52	.606
Hedged*CRT	0.03	0.03	1.02	.308
Detailed*CRT	−0.02	0.03	−0.68	.498
Hedged*Detailed*CRT	−0.01	0.03	−0.29	.769

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,587.

Table 6. General linear model predicting acceptance of information from correction style conditions and AOT as potential moderator.

	Estimate	Standard Error	*t*	*p*
Intercept	2.53	0.03	84.81	< .001***
Hedged	0.05	0.03	1.83	.068
Detailed Explanation	−0.02	0.03	−0.58	.563
AOT Score	0.21	0.03	7.00	< .001***
Hedged*Detailed	0.002	0.03	0.08	.935
Hedged* AOT	0.03	0.03	0.93	.353
Detailed* AOT	0.003	0.03	0.09	.930
Hedged*Detailed* AOT	−0.01	0.03	−0.19	.854

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$. Df = 1,501.

do not have a substantial impact on how corrections of misinformation on social media are received. These findings are consistent with recent research on correction wording and tone, which found that correction tone did not substantially affect misperceptions (see Bode et al., 2020). Our current findings extend this research in several keyways. First, we demonstrate that correction strength (roughly analogous to tone) does not significantly affect engagement with corrections of political misinformation, whereas prior work has looked at apolitical misinformation (Bode et al., 2020). Second, we also show that differences in correction strength do not impact engagement or belief updating by the user who shared the corrected misinformation themselves. This in contrast with previous work, which has instead assessed the effect of corrections on third-party viewers (observational correction, i.e., how third-party users on social media are affected by corrections; see Vraga & Bode, 2017). Third, we also show that manipulating the explanatory depth of the correction also has a minimal effect on engagement with the corrective message.

Our research also extends previous research on cognitive style and misinformation, which has found that people who are more reflective are less likely to believe false news headlines (Bronstein, Pennycook, Bear, Rand, & Cannon, 2019; Pennycook & Rand, 2019b, 2020), and that deliberation causally reduces belief in false claims (Bago, Rand, & Pennycook, 2020)—regardless of their partisan alignment. Here we examine the relationship between cognitive style on the response to corrections (rather than the perceptions of the misinformation itself). We found that analytic thinking and actively open-minded thinking (as assessed by CRT and AOT scales) predicted increased acceptance of corrective information. This willingness to update one's beliefs in the face of corrective information may help to explain why more reflective individuals have more accurate beliefs. Importantly, our results also suggest that analytic and actively open-minded thinking relate to increased acceptance of corrective information regardless of correction style.

Finally, our findings suggest that attempts at misinformation correction are not doomed to simply further promote dark engagement and incite comment section 'flame wars' (Flame war, n.d.). Indeed, self-reported belief updating was positive on average (M = 3.54), and average belief updating in reply texts as scored by Amazon Mechanical Turk raters (M = 2.63) was greater than the individual averages of all forms of resisting information (max: $M_{\text{Counter-arguing}}$ = 2.60). Thus, in line with previous research (e.g., Bode & Vraga, 2018), social media may not only serve as a medium for misinformation—online platforms may also enable and encourage user-generated corrections which, regardless of strength or explanatory depth, may be effective at combatting misinformation.

4.1. Limitations

The current research has several notable limitations. First, while we use a sample that is quota-matched to the American national distribution on age, gender, ethnicity, and geographic region, our findings may not generalize to other populations. Further research examining different countries and cultures, as well as underrepresented populations, is an important direction for future work.

Second, participants were not on their actual social media platforms, did not share fake news articles on their social media platforms, and knew that the correction they received was from a fictional account. Therefore, it is critical to test how the results of the current study generalize to more ecologically valid settings. Further research should examine the impact of manipulating corrective messages via a field experiment on a social media platform such as Twitter or Facebook.

Third, our study employs only one possible manipulation of hedging, and one possible manipulation of explanatory depth. Thus, it is plausible that other formulations of hedging or explanatory depth may yield differential engagement with corrective messages. For instance, our hedged message may be overly uncertain and perhaps more polite than other possible ways to hedge (e.g., "I'm not sure" vs. "It could be the case"). Thus, more certain and less polite hedged corrections may elicit greater engagement than the hedging manipulation we utilized. Furthermore, we definitionally manipulated explanatory depth by utilizing one condition in which the explanation was a simple negation, and the other condition included generic details about the source of the misinformation (i.e., "created by a website that purposefully makes false stories"). Given that perceptions of informativeness and unhelpfulness did not differ based on explanatory depth condition, it may be the case that either more detailed or more specific explanations may also lead to higher or lower levels of engagement with the corrective message. Future research may explore these possibilities in greater depth.

Fourth, many of our null results were not that precisely estimated. Thus, our findings should not be interpreted as evidence of no difference between correction conditions. Rather, our minimal and null results should be interpreted as a lack of evidence suggesting correction style does affect correction engagement—and, given our pre-registered prior hypotheses regarding likely differences in correction outcomes based on prior research, this lack of evidence was both surprising and complements recent research also indicating that correction style does not substantially impact correction engagement. Nonetheless, our minimal and null results should be interpreted with caution—we do not claim to find evidence of no effect of correction style on responses to misinformation, but rather present our results suggesting that our experiment yielded an absence of any evidence showing an effect of correction style.

5. Conclusions

In sum, we do not find evidence that hedging corrections of misinformation or providing increased explanatory depth in corrections of misinformation had a meaningful impact on engagement with corrective messages on social media. Although we found differences in how these messages were perceived in terms of aggressiveness or politeness, we did not find any substantial difference in likelihood of replying, overall acceptance of corrective information, or overall resistance towards corrective information. Our results also suggest that more analytic individuals, and more actively open-minded individuals, are more likely to accept corrective information, irrespective of correction strength or explanatory depth. Ultimately, our current study suggests that corrective messages, regardless of precise style or wording, may nonetheless be used as a source of positive engagement and communication on social media in order to combat dark participation.

Acknowledgments

The authors thank Antonio Arechar for assistance with data collection, and gratefully acknowledge funding from the William and Flora Hewlett Foundation, the Ethics and Governance of Artificial Intelligence Initiative, and MIT Libraries. This material is also based on work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 174530.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

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

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14). https://doi.org/10.1126/sciadv.aay3539

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613.

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

Bode, L., Vraga, E. K., & Tully, M. (2020, June 11). Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review*. Retrieved from https://misinforeview.hks.harvard.edu/article/do-the-right-thing-tone-may-not-affect-correction-of-misinformation-on-social-media

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117.

Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

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*, *42*, 1073–1095.

Clement, J. (2020, July 15). Number of global social network users 2017–2025. *Statista*. Retrieved from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/#:∼:text=Social%20media%20usage%20is%20one,almost%204.41%20billion%20in%202025.&text=Social%20network%20penetration%20is%20constantly,2020%20stood%20at%2049%20percent

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, *6*(1). https://doi.org/10.1177/2053168018822174

Epstein, Z., Pennycook, G., & Rand, D. (2020). Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–11). New York, NY: Association for Computing Machinery.

Flame war. (n.d.). In *Dictionary.com*. Retrieved from https://www.dictionary.com/browse/flame-war

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25–42.

Frenkel, S., Alba, D., & Zhong, R. (2020, March 8). Surge of virus misinformation stumps Facebook and Twitter. *The New York Times*. Retrieved from https://www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., & Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 324–332). New York, NY: Association for Computing Machinery.

Lakoff, G. (1975). Hedges: A study in meaning criteria and the logic of fuzzy concepts. In D. Hockney, W. Harper, & B. Freed (Eds.), *Contemporary research in philosophical logic and linguistic semantics* (pp. 221–271). Dordrecht: Springer.

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

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.

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.

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *Plos One*, *15*(2). https://doi.org/10.1371/journal.pone.0228882

Nyhan, B., & Reifler, J. (2012). *Misinformation and fact-checking: Research findings from social science*. New York, NY: New America Foundation.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4921–5484. https://doi.org/10.1287/mnsc.2019.3478

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (in press). Shifting attention to accuracy reduces online misinformation. *Nature*.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting Covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526.

Pennycook, G., & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200.

Prasad, M., Perrin, A. J., Bezila, K., Hoffman, S. G., Kindleberger, K., Manturuk, K., & Powers, A. S. (2009). "There must be a reason": Osama, Saddam, and inferred justification. *Sociological Inquiry*, *79*(2), 142–162.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931–2937). Copenhagen: Association for Computational Linguistics.

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10.

Shearer, E., & Matsa, K. E. (2018, September 10). News use across social media platforms 2018. *Pew Research Center*. Retrieved from https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*, 423–428.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*, 225–247.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web* (pp. 613–624). Republic and Canton of Geneva: International World Wide Web Conference Committee (IW3C2).

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*, 99–113.

Treventhan, R. (2017). Intraclass correlation coefficients: Clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, *17*(2), 127–143.

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. (2018). I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, *21*(10), 1337–1353.

About the Authors

Cameron Martel is a Graduate Student at Sloan School of Management, Massachusetts Institute of Technology. His research explores the cognitive processes underlying the belief and spread of misinformation. He is also broadly interested in social and political behavior online, particularly on social media platforms. In his research, he uses a variety of approaches from cognitive and social psychology, behavioral economics, and computational social science. He received his BS in cognitive science from Yale University. His work is supported by the National Science Foundation Graduate Research Fellowship.

Mohsen Mosleh is a Lecturer (Assistant Professor) at the Science, Innovation, Technology, and Entrepreneurship Department, University of Exeter Business School. Mohsen was a Postdoctoral Fellow at the MIT Sloan School of Management as well as the Department of Psychology at Yale University. Prior to his postdoctoral studies, Mohsen received his PhD from Stevens Institute of Technology in Systems Engineering with a minor in data science. Mohsen's research interests lie at the intersection of computational/data science and cognitive/social science. In particular, he studies how information and misinformation spread on social media, collective decision-making, and cooperation.

David G. Rand is the Erwin H. Schell Professor and Associate Professor of Management Science and Brain and Cognitive Sciences at MIT. His research combines behavioral experiments and online/field studies with mathematical/computational models to understand human decision-making. David focuses on illuminating why people believe and share misinformation and 'fake news,' understanding political psychology and polarization, and promoting human cooperation. He was named Poynter Institute Fact-Checking Researcher of the year in 2017, and received the 2020 FABBS Early Career Impact Award from the Society for Judgment and Decision Making.