The Effects of Message Order and Debiasing Information in Misinformation Correction

YUE DAI[1]
WENTING YU
FEI SHEN
City University of Hong Kong, Hong Kong

Misinformation continues to influence inferences even after being discredited, making it extremely difficult to completely erase its detrimental effects. With a two-wave online experiment, this research tested how the effectiveness of misinformation correction is influenced by (1) whether correction is presented before or after misinformation and (2) whether correction is accompanied by a message that enhances the coherence between misinformation and correction message. The results showed that a correction was most effective when it was delivered after the misinformation and with a debiasing message. These effects persisted at least one week after the initial exposure to the correction. The results were consistent with the Knowledge Revision Components (KReC) framework and the schemata-plus-tag model of negation comprehension. The findings also provided a comprehension-based explanation to previous findings from meta-analysis regarding the order of presentation of misinformation and corrective messages. Practical implications for misinformation correction practices are discussed.

Keywords: misinformation correction, primacy effect, recency effect, coherence, memory, inoculation

Misinformation and fake news have become a global issue, given the quick expansion of social media use. Nations around the world are taking action to fight against misinformation, such as passing fake news laws, improving citizen media literacy, and regulating social media (Reuters, 2019). In particular, social media have been criticized for facilitating the spread of untruthful information in the domains of politics (Lewandowsky, Ecker, & Cook, 2017), health, and science (Bessi et al., 2015).

Yue Dai: nancy.dai@cityu.edu.hk
Wenting Yu: wentingyu3-c@my.cityu.edu.hk
Fei Shen: feishen@cityu.edu.hk
Date submitted: 2020-05-04

[1] This research is supported by the Faculty Start-up Fund (Project No. 7200638) and the Strategic Research Grant (Project No. 7005330) at City University of Hong Kong.

Among the many measures taken to curb the negative consequences of untruthful online information, using verified information to correct misinformation is a direct solution to the problem (Walter & Murphy, 2018). A burgeoning body of research on misinformation in recent years has identified factors that influence the effectiveness of fact-checkers and misinformation-correction strategies in general (Walter, Cohen, Holbert, & Morag, 2020). For example, past research has focused on the format and the delivery source of fact-checkers (e.g., Ecker, O'Reilly, Reid, & Chang, 2020; van der Meer & Jin, 2020), their congruence with the message recipients' prior attitude (e.g., Ecker & Ang, 2019; Hameleers & van der Meer, 2020), their effectiveness when combined with other interventions (Hameleers, 2020), how they spread in new media contexts (Margolin, Hannak, & Weber, 2018), and the potential mechanisms through which people process misinformation and its correction (e.g., Crozier & Strange, 2019; Jang, Lee, & Shin, 2019).

Despite the consensus that fact-checking and correction are desirable and that free flow of information facilitates the spread of truth, there is still much to be understood about effectively correcting misinformation. Although a substantial and burgeoning body of research has been dedicated to effective misinformation correction (Seifert, 2002), a less optimistic finding is that the influence of misinformation will persist even when correction is provided (Ecker, Lewandowsky, & Tang, 2010). Evidence for the continued effect of misinformation was discovered in multiple empirical tests of misinformation correction strategies (Ecker, Lewandowsky, Cheung, & Maybery, 2015; Gordon, Brooks, Quadflieg, Ecker, & Lewandowsky, 2017; Johnson & Seifert, 1994; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Thorson, 2016). In all these studies, even though correction messages ameliorated the detrimental effects of misinformation, the retracted information continued to be salient in participants' memory and thereby influenced their inferences. It seems, then, that one key to improving the effect of misinformation correction messages is to understand what conditions facilitate information recipients' cognitive reliance on the correct information and relinquish the information that is proved inaccurate.

In this study, we explored two factors that might influence the effectiveness of misinformation correction. The first factor we focused on was message order. That is, is it more effective to present the correction information before or after the misinformation? The order of information presentation may have an effect on people's memory and comprehension, especially when two pieces of information are in conflict with each other, as in the case of misinformation correction. But few studies have systematically manipulated the order of misinformation and correction message. Based on models of knowledge update (Kendeou & O'Brien, 2014) and the literature on negation comprehension (Mayo, Schul, & Burnstein, 2004), we propose hypotheses regarding the relative effectiveness of presenting corrective messages before versus after the misinformation.

In terms of message content, we explored the function of a debiasing message (Lewandowsky et al., 2012) as a way of attenuating people's continued reliance on misinformation even after it is retracted. Previous research revealed that providing a plausible causal alternative, rather than simply retracting misinformation, effectively decreased people's continued reliance on misinformation (Johnson & Seifert, 1994). Prior research also revealed that coherence *within* the information and coherence *between* people's prior attitude and the information enhance their information processing and comprehension (Hardisty, Johnson, & Weber, 2010; Lewandowsky et al., 2012). When comprehending conflicting information, individuals will generate causal inferences to enhance the coherence of the information in order to ease the information processing (Blanc, Kendeou, van den Broek, & Brouillet, 2008). Building on the prior findings, this research investigates whether providing a debiasing message that reconciles the conflict between two alternative accounts of an event enhances the outcome of misinformation correction.

Finally, we explored the persistence of the mentioned effects of presentation order and debiasing message by measuring participants' memory of and attitudes toward the relevant issue, both immediately after the exposure and one week later.

The Continued Influence of Misinformation

Although misinformation can be used to label a wide range of information (Molina, Sundar, Le, & Lee, 2019), the present research defines misinformation broadly without limiting the scope to a specific taxonomy. Following definitions that were established in prior research (e.g., Lewandowsky et al., 2012), the current research defines misinformation as any piece of information that is retracted or corrected for its untruthful representation of an event.

Abundant empirical evidence has shown that despite efforts to correct for information that is untruthful, discredited information continues to exert its influence on people's perception of the relevant event (Ecker et al., 2015; Gordon et al., 2017; Johnson & Seifert, 1994; Lewandowsky et al., 2012; Thorson, 2016). One study, for example, compared different warning strategies to correct misinformation in a fictitious story about a minibus accident. The study found that although retraction messages worked to correct participants' memory of the story, participants continued to remember information that was explicitly stated to be inaccurate in a subsequent recall task after they read the stimuli (Ecker et al., 2010).

Processes and Mechanisms of Knowledge Revision

At a theoretical level, misinformation correction is a process of updating existing knowledge about an event. When encountering misinformation corrections, an individual encodes new information about a subject that often conflicts with what he or she already knows about the subject (Kendeou & O'Brien, 2014).

The Knowledge Revision Components (KReC) framework describes the processes and the mechanisms of knowledge revision. According to KReC, knowledge revision is an incremental and slow process. The presence of new and conflicting information about a subject will reinstate the previous information about a subject. This will trigger a bottom-up process in which new information about the subject makes contact with the previous knowledge and reactivates it, such that both pieces of information will appear in working memory at the same time. Knowledge revision eventually occurs as new information about the subject is encoded and integrated with the outdated information (Kendeou & O'Brien, 2014). Noteworthy is that, based on the description of KReC, the reactivation process is essential for the eventual integration of the previous and the new information to occur. If this is true, one should be able to enhance the effectiveness of misinformation correction by facilitating the reactivation of previous information when delivering the correct information.

Information Presentation Order

A person seeking information online may encounter misleading information about an issue or event and subsequently be exposed to debunking messages about it. Alternatively, a person may read corrective information about an issue on a fact-check website and subsequently learn about the misinformation that the correction addresses. Both formats are present in well-established fact-check websites. For example, Snopes.com presents the title of each fact-check in a question format. On clicking on a title of interest, users will read the misleading information, followed by a verdict of its truthfulness and detailed corrective messages. In a free information-searching context, such as Twitter, however, users may read a debunking message with the original misleading information tagged onto it, in which case a reader may be exposed to the misinformation after the correction.

Although real-life scenarios reflect different presentation orders of misinformation and corrective information, the relative presentation order between the two has rarely been directly examined in the research on misinformation. A recent meta-analysis on 65 studies on misinformation correction (total *N* = 23,604), however, suggested that information presentation order might be a factor that influences the outcome of misinformation correction. The study revealed that refuting the misinformation after it was released was more effective (in terms of the effect size) than warning individuals against potential erroneous information before they were exposed to the misinformation (Walter & Murphy, 2018).

To understand the theoretical reason for the effect of information presentation order, one needs first to understand how individuals cognitively process negations. There are two theoretical accounts of negation processing: the fusion model and "schema-plus-tag" model (Mayo et al., 2004). In the fusion model, information perceivers activate an affirmative schemata that is opposite to the meaning of the negated message and encode that schemata. For example, in processing the message "John is not a hard-working student," one would imagine lazy behaviors from John in encoding the message. In the schema-plus-tag model of negation encoding, however, the information receiver does not activate an opposite schema, but rather encodes the message in the format of "A is Not(X)." In this example, the message would be coded as "John is *not* hard-working" under this model (Rapp & Braasch, 2014). A key condition that determines which processing mode is used is whether a clear affirmative schemata that is opposite to the negated message is readily available, which is likely only if the negated information has a definitive polar opposite; otherwise, people are likely to process negation according to the schemata-plus-tag model (Mayo et al., 2004; Rapp & Braasch, 2014), which, we argue, tends to be the case in most of the misinformation correction scenarios.

It is worth noting that processing negation under the schemata-plus-tag model requires the coactivation of negation and the information that is being negated (Mayo et al., 2004). This contention is supported by empirical findings showing that people may misremember the negated information as the true version of the story. In a series of experiments, Maciuszek and Polczyk (2017) asked participants to listen to a description of a house. Some objects were mentioned, some negated, and some not mentioned. The results showed that many participants misremembered the negated objects as having been mentioned, an effect they termed "negation-related false memories (NRFM)" (Maciuszek & Polczyk, 2017, p. 3). This phenomenon was replicated on a sample of 5- and 6-year-olds and using stimuli that described actions. These results showed that outdated information was activated in the process of comprehending negations.

If the activation of the negated information is necessary for knowledge update as KReC predicts, it is possible that presenting the negation *after* the original information is more effective than presenting it *before*. When negation is delivered after the misinformation, a reader would need to activate the previous outdated information to comprehend the negation. But when negation is encountered by an information seeker before the misinformation, a message reader may not activate the negation when he or she reaches the information that is meant to be corrected because it is not a part of the comprehension step. As a result of the lack of activation, it could weaken the knowledge update process. Therefore,

H1: In misinformation correction, people will remember the correct information more when they are exposed to corrective messages after misinformation rather than before misinformation.

Cognitive Coherence

In addition to the relative presentation order of the misinformation and its correction, another factor that can influence the effectiveness of the correction is the coherence between the correction and the misinformation. Previous research on information comprehension and retention has discovered that people comprehend coherent information more effectively. Individuals engage in two types of comprehension processes simultaneously when comprehending information—an automatic memory-based process in which they take in the information passively, and a constructive process in which they actively connect pieces of incoming information with logical reasoning (van den Broek, Risden, & Husebye-Hartmann, 1995). These processes, according to van den Broek and colleagues (1995), operate by the principle of coherence: If readers feel that enough coherence of the story has been yielded from the automatic memory-based process of comprehension, they will not activate the constructionist process; otherwise, a strategic construction process will activate until the readers find a satisfactory level of coherence in the story or abandon the desire to comprehend the story further. Similarly, Lewandowsky and associates (2012) note that coherent information is more resistant to change than incoherent stories. According to them, a piece of information will be accepted as true and will be retained longer in memory if it fits with the rest of a story and lends a sense of coherence to an individual.

Empirical research has lent support to the idea that people actively seek logical connections in information comprehension (Blanc et al., 2008; Devine & Ostrom, 1985). Johnson and Seifert (1994) conducted an experiment in which participants were shown different versions of a story about an investigation of a warehouse fire. The study discovered that people continued to make inferences involving discredited information when it provided a plausible explanation to the fire in the story (i.e., a short circuit near a closet reportedly contained volatile materials). But no continued influence of misinformation was found when the same information was primed incidentally with an intervention task. In other words, participants remembered the misinformation longer and used it for subsequent inferences when the misinformation could be logically connected to other information in the story, despite knowing that the information was not accurate. These results suggest that information readers do actively seek internal coherence in information processing and comprehension.

In misinformation retraction, the correction message often negates the legitimacy of the misinformation first and presents an alternative subsequently (van den Broek et al., 1995). Previous research has encouraged the provision of an alternative account of the issue in misinformation correction as it fills the cognitive gap left from the retraction of a previous account (Johnson & Seifert, 1994). One characteristic of this misinformation correction strategy, however, is that it is likely to contain contradictions, given that the correction message and the original misinformation inevitably offer two alternative accounts of the same event. If coherent information is better memorized and is more resistant to changes (Lewandowsky et al., 2012), it should enhance the effectiveness of misinformation correction when a correction message contains information that helps people resolve the conflicts between the misinformation and the correct alternative. One example of such information is a message that explains *why* the misinformation exists in the first place (hereafter referred to as debiasing message; Lewandowsky et al., 2012). Therefore,

H2: The presence of a debiasing message that reconciles the conflict between the information and its correction increases individuals' recall of the correction message.

So far, H1 and H2 have articulated parallel effects of the order of presentation and the presence of debiasing messages on people's recall of and attitude toward an issue, respectively. This causal chain could be further extended to an individual's attitude toward an issue to which the misinformation and the correction pertain. Prior research has indicated that people's attitude toward an issue or event is influenced by their recall of the information related to that event. For instance, in a series of experiments on storytelling, memory, and impressions, McGregor and Holmes (1999) asked participants to read stories about a relational conflict between two characters in which both parties could be blamed, depending on the reader's interpretation of the story. The results revealed that participants' biased recall of the details of the conflict significantly influenced their assignment of the blame to the characters. More recently, results from a meta-analysis revealed that changes in attitude and behavior intentions were contingent on the recall of persuasive messages: the better the recall, the greater the changes (Blondé & Girandola, 2016). The following two hypotheses (H3 and H4) describe how recall of the correction message mediates the effects of message order and debiasing message on individuals' attitude toward the relevant issue (Figure 1).

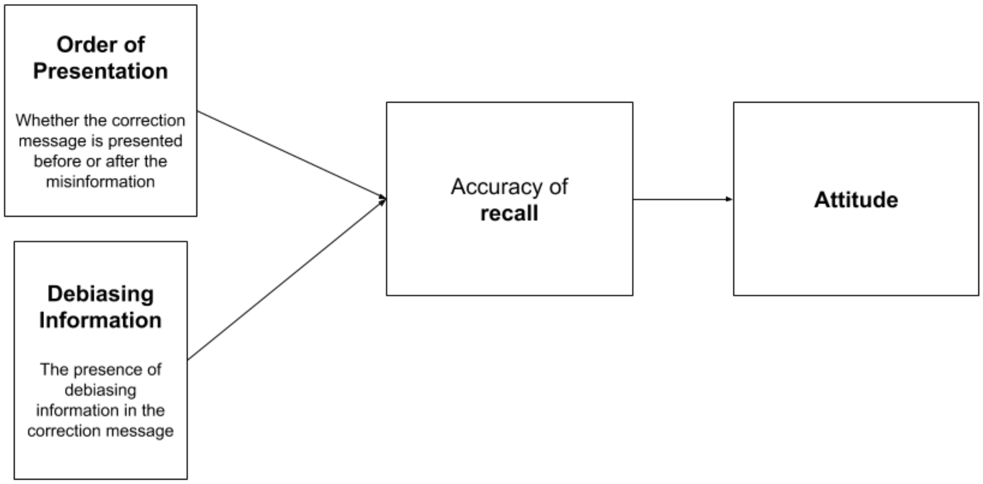
Order of Presentation
Whether the correction message is presented before or after the misinformation
Debiasing Information
The presence of debiasing information in the correction message
Accuracy of recall
Attitude

Figure 1. A conceptual model of the factors examined in the study.

H3: *An individual's recall of the correction message to misinformation significantly influences the individual's attitude toward the relevant issue.*

H4: *Individuals' recall of the correction message mediates the effects of presentation order and debiasing message on people's attitude toward the relevant issue.*

Method

Design and Stimuli

We conducted a two-wave online experiment that featured a mixed-factorial design. Time was a within-subject factor. The three between-subject factors were (1) labeling (whether a message was labeled as misinformation or accurate information), (2) order (whether the correction appeared before or after the misinformation), and (3) debiasing information (present vs. absent). Participants were randomly assigned to one of eight experimental conditions and completed the same questions twice with a week in between.

Before reading the experimental stimuli, participants were told that they would read a story about a car accident and then answer some questions about it. They were then presented with a fictitious story written in plain text without being embedded in any interface. The source of the story was also not identified. Using a fictitious story made it much easier to create two accounts of the same story that were comparable in their memorability so that any ordering effect discovered was not confounded with message quality. A fictious story also helped to avoid any overriding effects of participants' preexisting attitudes or their previous exposure to the story, making it possible to switch the labels of the key messages in the story.

Following Ecker and colleagues (2010), our stimuli featured a fictitious traffic accident story. The key messages concerning the manipulation are presented in the online appendix.

All versions of the story offer two alternative accounts of why the accident happened, one blaming the accident on a passenger (sentence 4a in the online appendix) and the other on the driver (sentence 14b in the online appendix). Depending on the condition, one of these accounts was labeled as misinformation and the other as a correction. By altering the labels to the two accounts and measuring participants' attitude toward both the driver and the passenger, we built an internal replication within the design. The relative presentation order of the misinformation and the correction was also varied. In the correction-first conditions, participants read the true cause in paragraph 4 and the inaccurate cause in paragraph 14 of the story. In the misinformation-first conditions, paragraph 4 presented the inaccurate cause, whereas paragraph 14 presented the accurate one. Finally, depending on the condition, participants either read a debiasing message that explained what led to an inaccurate cause in the first place (paragraph 15), together with the correction information, or saw no such message.

It should be noted that in real life, people are more likely to read misinformation and its correction in the form of stand-alone articles on Twitter or Snopes.com, but our experimental inductions were realized through key messages in one article rather than in stand-alone articles. The decision was made to avoid having to repeat the filler messages in both articles and thereby unnecessarily emphasizing them over the key messages and biasing the results on recall.

Procedure

In the first wave of the study, participants read the stimuli article first and completed free recall questions about their memory of the accident. They also indicated their attitude toward the driver and the passenger in the story before answering the manipulation check questions and demographic questions. A week later, we invited participants to answer the same set of questions they did in Wave 1 without reading the article.

Pilot Study

Because one of the main dependent variables is participants' recall of the story, we conducted a pilot study to select two causes of the traffic accident that were equally plausible and memorable to participants to avoid any confounding effects caused by idiosyncratic features of the key messages.

We recruited 41 participants from the United States through Amazon's Mechanical Turk (MTurk), where collected samples are found to be more diverse than student samples (Buhrmester, Kwang, & Gosling, 2016). All participants received $1 in exchange for their participation. Participants first read the background story of a van accident and then rated the believability and memorability of six possible causes. The text length of each cause was held at around 30 words. We measured believability with five bipolar adjectives on a 7-point semantic differential scale: (1) unbelievable–believable, (2) unconvincing– convincing, (3) not plausible–plausible, (4) inauthentic–authentic, and (5) not credible–credible. The measurement of memorability consisted of four statements: "I think the cause is easy to remember," "The information would register easily in people's mind," "I would have no difficulty recalling the information described in the message after reading it shortly," and "I would have no difficulty recalling details in the message several days after reading it." Participants were asked to rate these items on a 7-point Likert scale (1 = *strongly disagree*; 7 = *strongly agree*).

We ran two analyses of variance (ANOVAs) to compare different potential causes on their believability and memorability (Table 1).

Table 1. Descriptive Statistics of Participants' Perceived Believability and Memorability of the Six Stimulus Messages From the Pilot Study.

		Believability			Memorability		
		M	*SD*	*a*	*M*	*SD*	*a*
1.	The accident happened because a speeding car kept swinging around the van. Three teenagers in the speeding car—two females and a male—were arrested for suspected dangerous driving.	4.42^{a}	1.77	.99	4.99	1.38	.94
2.	The accident happened because an off-leash dog suddenly rushed to the road. The driver made a few sharp turns to avoid the dog but eventually bumped into the steep embankment.	4.39^{be}	1.53	.96	5.24	1.14	.87
3.	The accident happened because the driver was exhausted. A passenger on board said the driver had been driving continuously for 18 hours and that he was distracted when the accident happened.	5.73^{abcd}	1.10	.96	5.52^{f}	1.13	.95
4.	The accident happened because the driver did not signal properly before making a turn. A car from the next lane failed to respond in time and hit the van from behind (selected message).	4.97^{c}	1.44	.97	4.80^{fg}	1.37	.93
5.	The accident happened because the driver was distracted by a standing passenger on board. In trying to warn the passenger, the driver failed to notice a stop sign and crashed into the embankment (selected message).	4.88^{d}	1.26	.95	5.04^{h}	1.02	.77
6.	The accident happened because of drunk driving. The driver attended a friend's birthday party two hours before he picked up the passengers. He is now charged with suspected driving under the influence.	5.39^{e}	1.43	.97	5.66^{gh}	1.01	.90

Note. Different superscripts indicate significant differences in pairwise comparisons.

Based on the results, three pairs of causes did not differ significantly in either memorability or believability: Cause 1 and 5, Cause 1 and 2, and Cause 4 and 5. We then ran an equivalence test following Weber and Popova (2012) to identify the most equivalent pair. We found that Causes 4 and 5 were significantly equivalent at the moderate effect size on believability, Δ =.03, *t*(41) = .48, *df* = 40, *p* = .60 (two-tailed), and memorability, Δ = .03, $t(41) = -1.29$, $df = 40$, $p = .86$ (two-tailed). That is, the difference between them was not enough to constitute a moderate-sized effect. Cause 4 (sentence 14b in the online appendix) and Cause 5 (sentence 4a in the online appendix) were hence selected for the main study.

Participants

The sample for the main study was also recruited from the United States through MTurk. Wave 1 of the data collection took place August 16–18, 2019. A total of 350 participants completed the study in exchange for $1. A week later, about 84% of them participated in the second wave of the data collection and received another $1. To match participants' answers in both waves while keeping them anonymous in the data set, we asked them to create a unique code that featured a combination of their mothers' maiden name and their birthdays. Unfortunately, some participants could not recall their codes in the second wave of the data collection. After excluding the participants who could not be matched, we obtained a final sample of $N = 271$ individuals.

Participants ranged in age from 20 to 73 ($M = 36.76$, SD = 11. 34). Males constituted 57.2% of the sample. The three largest racial groups of the sample were Caucasian (71.2%), African American (10.7%), and Asian (8.9%). Each experimental condition had 31–38 participants.

Measurement Scales

Because participants' recall and attitude are highly contextualized, we created a series of original measurement items for recall and attitude. In the following description, we use the subscripts 1 and 2 to denote the relevant statistics from Wave 1 and Wave 2 data.

Attitude Toward the Responsible Party

Participants were asked to rate how they felt about the van driver and passenger's behavior, respectively, on a 7-point semantic differential scale featuring six bipolar adjectives: bad/good, negative/positive, unacceptable/acceptable, unfavorable/favorable, foolish/wise, and dishonest/honest. The mean scores of these items formed the indices of attitude toward the van driver ($M_1 = 3.16$, $SD_1 = 1.43$, $\alpha_1 = .84$; $M_2 = 3.31$, $SD_2 = 1.53$, $\alpha_2 = .97$) and attitude toward the standing passenger ($M_1 = 2.96$, $SD_1 = 1.66$, $\alpha_1 = .98$; $M_2 = 2.98$, $SD_2 = 1.65$, $\alpha_2 = .98$), respectively. To represent the extent to which the participants blamed the correct person (driver vs. passenger), we built an index attitude score based on the conditions they were assigned to. For the driver-at-fault conditions, we subtracted participants' attitude toward the driver from their attitude toward the passenger. For the passenger-at-fault conditions, we subtracted participants' attitude toward the passenger from their attitude toward the driver. Therefore, a positive score on the attitude measure indicates that the participant favored the innocent party more than the responsible party. In both waves of the data, participants showed a more positive attitude toward the innocent party in ways that were consistent with the condition to which they were assigned ($M_1 = 1.30$, $SD_1 = 1.98$; $M_2 = 1.381$, $SD_2 = 2.15$).

Fact Recall Accuracy

Following Ecker and associates (2010), we used open-ended questions to test participants' memory on the stimulus stories. Immediately after reading the stimulus story, participants answered three open-ended questions regarding the differential aspects of the two versions of the stimuli story: (1) the real cause of the accident, (2) the earlier findings of the accident, and (3) the responsible person. In the subsequent coding of participants' answers, we coded completely correct answers as 1, partially correct answers (i.e., recalling the faulty party correctly with inaccurate details) as 0.5, and completely incorrect or irrelevant answers (i.e., "I don't know") as 0. Two trained coders scored the open-ended questions from 32 randomly selected participants (four from each condition) for an interrater reliability check. The interrater reliability of these three questions suggested a high agreement between the two coders, with Cohen's kappa ranging from .87 to .95 for both waves. Then the coders continued to code all the answers. We added up participants' scores on all three open-ended recall questions to form an index of memory accuracy that ranged from 0 to 3. The reliability of the three questions was .60 at Wave 1 and .72 at Wave 2. Not surprisingly, participants' fact recall of the accident at Wave 1 (*M* = 2.26, *SD* = .96) was more accurate than that at Wave 2 (*M* = 1.81, *SD* = 1.21).

Manipulation Checks

Three questions assessed whether participants had noticed our manipulations on order, labeling, and debiasing message, respectively. For order, participants were asked, "In the story, you read about two causes of the accident—an inaccurate cause concluded in an earlier police statement, and a true cause concluded in a more recent police statement. Do you remember in what order you read about the two causes?" They were given three choices for the question: "I read about the inaccurate cause from an earlier investigation and then the true cause from a more recent investigation," "I read about the true cause from a recent investigation first and then the inaccurate cause from an earlier investigation," and "I can't remember the order." For labeling, participants were asked, "Do you remember which message below describes the true cause of the accident in the story you read?" They were asked to choose among the following three choices: "The accident happened because the driver was distracted by a standing passenger on board. In trying to warn the passenger, the driver failed to notice a stop sign and crashed into the embankment," "The accident happened because the driver did not signify properly before making a turn. A car from the next lane failed to respond in time and hit the van from behind," and "I can't remember clearly." Finally, for debiasing message, the induction question read, "In the story you read, did you see any information about what has led to the inaccurate conclusion of the cause of the accident in an earlier investigation?" Participants were asked to choose from "yes," "no," and "I'm not sure." Each question was recoded as a binary variable. The correct answers were coded as 1, and the incorrect answers (including the "I don't know" options) were coded as 0.

Results

Manipulation Checks

To test whether participants who were assigned to read each of the two versions of the story correctly recognized the true cause of the accident in the version they saw, a chi-square test examined the association between the manipulation of the two versions of the story (driver's fault vs. passenger's fault) and participants' recall of the true cause of the accident in the manipulation check question. Results showed that the two variables were significantly associated at both Wave 1, $\chi^2(1, N = 271) = 146.61$, $p < .001$, Cramer's $V = .74$, and Wave 2, $\chi^2(1, N = 270) = 108.09$, $p < .001$, Cramer's $V = .63$. The manipulation of labeling was hence deemed successful.

To test the effectiveness of the order manipulation, we conducted another chi-square test to examine the association between the presented order of correction and participants' recall of the order in which the misinformation and the correction messages were presented. These two variables were significantly associated at both Wave 1, $\chi^2(1, N = 271) = 115.74$, $p < .001$, Cramer's $V = .65$, and Wave 2, $\chi^2(1, N = 271) = 115.74$, $p < .001$, Cramer's $V = .65$. Although the manipulation was effective in the sense of statistical significance, it should be noted that 48 of 271 participants remembered that the misinformation was presented first, while the stimuli presented the correction information first. It is possible that this occurred because the question options described the misinformation as being concluded "from an earlier investigation," so participants assumed that information from the earlier report should be presented first in the article.

To test the effectiveness of the manipulation of debiasing message, a chi-square test examined the association between whether a debiasing message was shown and participant's recall of whether he or she read about why the previous investigation led to a wrong conclusion. The chi-square test revealed a significant association between these two variables at both Wave 1, $\chi^2(1, N = 271) = 110.88$, $p < .001$, Cramer's $V = .64$, and Wave 2, $\chi^2(1, N = 271) = 110.88$, $p < .001$, Cramer's $V = .64$. The manipulation of the debiasing message was hence deemed successful.

Hypothesis Testing

Because participants' recall was measured at two time points, we ran a repeated-measures ANOVA to test the effects of the experimental inductions and to see whether the results differed across time. Participants' recall accuracy of the story acted as the dependent variable for the ANOVA model. In the ANOVA model, time acted as a within-subject factor, given that every participant was measured at two time points. This ANOVA model tested Hypotheses 1 and 2.

Hypothesis 1 posits that the order in which a correction message is presented relative to the misinformation influences individuals' recall of the correction message. Results of the ANOVA test revealed a significant effect of message order on participants' recall accuracy, $F(1, 262) = 25.01$, $p < .001$, partial $\eta^2 = .087$. Compared with presenting the correction *before* the misinformation ($M = 1.77$, $SD = .04$), a correction message *after* the misinformation ($M = 2.29$, $SD = .07$) led to more accurate recall of the story

($p < .001$). The lack of a significant time-by-order interaction indicated that the effect of message order did not change over time, $F(1, 262) = 0.00$, $p = .95$.

Hypothesis 2 posits that the presence of a debiasing message that reconciles the conflict between the information and its correction increases individuals' recall of the correction message. The ANOVA results revealed a significant effect of debiasing message, $F(1, 262) = 12.84$, $p < .001$, partial $\eta^2 = .047$. Participants who read a debiasing message ($M = 2.21$, $SD = .07$) scored higher on memory accuracy ($p < .001$), compared with those who did not read a debiasing message ($M = 1.84$, $SD = .07$). The effect of debiasing did not change over time; there was no time-by-debiasing interaction, $F(1, 262) = 0.06$, $p = .80$. Noteworthy is that only two main effects of presentation order and debiasing message were discovered. The two factors did not interact to influence recall accuracy, $F(1, 262) = 1.31$, $p = .90$.

In addition to the mentioned results, the ANOVA model that tested H1 and H2 also revealed a significant effect of time, $F(1, 262) = 50.99$, $p < .001$, partial $\eta^2 = .163$. Not surprisingly, participants were less accurate in their recall of the event at Time 2 ($M = 1.81$, $SD = .07$) than at Time 1 ($M = 2.25$, $SD = .05$). Hypothesis 3 posits a positive correlation between fact recall and attitude. The analyses revealed a positive correlation between participants' recall accuracy and the difference in their attitude toward the driver and the passenger at both Time 1 ($r = .43$, $p < .001$) and Time 2 ($r = .54$, $p < .001$) of the measurement. As such, the data were consistent with H3.

Hypothesis 4 posits that fact recall mediates the effect of order and debiasing message on participants' attitude. The mediation model was tested with Hayes's (2013) PROCESS (Model 4). Because the model involves two exogenous variables (i.e., order of presentation and debiasing information), the analysis for the same model was run twice—each time using one variable as the exogenous variable and the other as a covariate—to obtain separate mediation indices for both exogenous variables in the model (Hayes, 2013). Figure 2 reports the individual path coefficients in the model. Overall, the results indicated that presenting the correction message after the misinformation and presenting it together with a debiasing message led to more accurate recall of the correction, which further influenced participants' attitude. The model explains a significant amount of the variance in the attitude at Time 1, $F(3, 262) = 26.67$, $p < .001$, $R^2 = .23$, and Time 2, $F(3, 262) = 40.87$, $p < .001$, $R^2 = .32$.

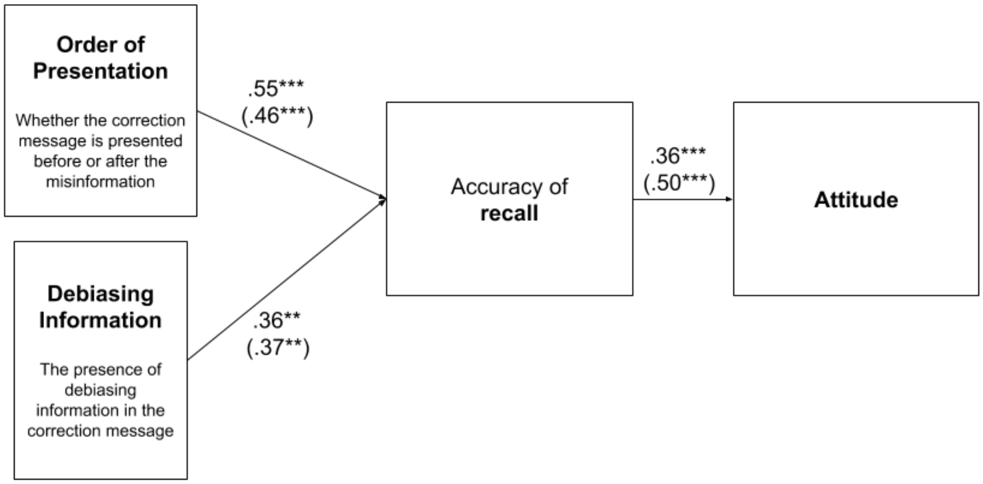
Order of Presentation
Whether the correction message is presented before or after the misinformation
Debiasing Information
The presence of debiasing information in the correction message
.55***
(.46***)
.36**
(.37**)
Accuracy of recall
.36***
(.50***)
Attitude

Figure 2. Path coefficients and statistical significance of moderated mediation analyses on attitude. Coefficients for analyses on Time 2 data appear in parentheses. All coefficients are unstandardized. **p < .01. ***p < .001.

The mediation index showed that the order of presenting misinformation and correction had a significant indirect effect (*b* = 0.39, *SE* = 0.10, 95% CI [0.21, 0.59]) through recall and a significant direct effect on attitude (*b* = 0.59, *SE* = 0.22*,* 95% CI [0.15, 1.03]) at Time 1. At Time 2, the order of correction and misinformation presented had a significant indirect effect (*b* = 0.49, *SE* = 0.14*,* 95% CI [0.24, 0.79]) on participants' attitude through recall, but no significant direct effect (*b* = 0.34, *SE* = 0.23, 95% CI [−0.11, 0.78]) was found. These results suggest that the accuracy of recall partially mediated the effect of order on attitude at Time 1, and it fully mediated the effect of order on attitude at Time 2.

As for whether recall mediated the effect of debiasing information on attitude, the analysis showed that both the estimated indirect effect (*b* = 0.26, *SE* = 0.09*,* 95% CI [0.09, 0.42]) and the estimated direct effect (*b* = 0.67, *SE* = 0.22*,* 95% CI [0.24, 1.10]) of debiasing message on attitude were significant at Time 1. Results at Time 2 showed a similar pattern, that the estimated indirect effect of debiasing on attitude was significant (*b* = 0.40, *SE* = 0.14*,* 95% CI [0.14, 0.68]), as was the estimated direct effect (*b* = 0.61, *SE* = 0.22*,* 95% CI [0.17, 1.05]). The results suggest that accuracy of recall fully mediated the effect of debiasing messages on attitude at Time 1 and Time 2.

Discussion

The present research was designed to test two factors that influence the effectiveness of misinformation correction on people's memory and, subsequently, on their attitude toward the relevant issue: (1) the relative presentation order of the misinformation and its correction message, and (2) whether the correction message was presented with a debiasing message that reconciled the conflict between the misinformation and the correction message.

Theoretical and Practical Implications

Although previous research has largely revealed the effectiveness of misinformation correction, it nonetheless showed that the effect of correction dissipates over time and that correction messages can never completely erase the erroneous impressions left by the misinformation (Ecker et al., 2010; Lewandowsky et al., 2012). Not surprisingly, we also observed that the accuracy of participants' memory of the event dissipated over time. Still, the results of our study offer two directions to enhance people's memory of correct information over time.

At a theoretical level, the results regarding the effect of information order were consistent with the proposition of the KReC framework (Kendeou & O'Brien, 2014) and the schemata-plus-tag model of negation comprehension (Mayo et al., 2004). Both models propose that for successful comprehension of negation and knowledge update to happen, both the negation and the negated information have to be coactivated in a person's mind. Specifically, the schema-plus-tag model of negation comprehension predicts that negation of information is more effectively comprehended when it is delivered after the information instead of before it because the former facilitates the coactivation of the inaccurate information and the negation tag. Supporting this contention, the results showed that presenting corrective messages *after* the misinformation influenced participants' recall of the event and shaped their attitudes to be more consistent with the truth more effectively than presenting the correction *before* the misinformation. This effect was replicated in the analysis on both the driver and the passenger in the story, demonstrating the robustness of the results.

In addition, these results provided a comprehension-based explanation to findings from earlier meta-analysis showing that correcting misinformation was more effective than inoculating people against it (Walter & Murphy, 2018). Although a forewarning and a retraction may differ on other characteristics, the results from this research suggest that, when other factors are held equal, the sheer order of information presentation may play a role in determining the effectiveness of misinformation correction. At a practical level, the findings regarding the effect of information presentation order have implications for misinformation correction practices. To contextualize fact-check messages, it is not uncommon for fact-check websites to present the original misinformation together with correction messages. The results of the study suggest that it may be beneficial to present corrective messages after rather than before the misinformation. Additionally, when disseminating debunking messages on social media, it may be worthwhile to first briefly summarize the event that the message seeks to debunk and then give the correction message; readers would not need to seek the original misinformation to understand the context, which, according to the results of the present study, leads to a less optimal outcome of misinformation correction.

Regarding the effect of debiasing information, the findings supported the literature on people's preference for coherent information (van den Broek et al., 1995). We discovered that participants' memory of the correction message, and, subsequently, the effectiveness of misinformation correction were enhanced by presenting a message that reconciled the conflicts between misinformation and the correction. That the enhancing effect of debiasing messages lasted for at least a week after the experiment supported the proposition that coherent information is resistant to change (Lewandowsky et al., 2012). These results suggest that when it is inevitable to expose people to misinformation and a potentially conflicting correction at the same time, one way to enhance information recipients' accurate memory is to explain what resulted in the misinformation in the first place.

With regard to the effect of time, the study discovered that participants' accurate recall of the story dissipated over time. Despite the natural deterioration of memory, the effects of presentation order and debiasing message were not found to be moderated by time. These results provided encouraging empirical evidence that presenting a correction after people's exposure to misinformation and presenting a correction with a debiasing message can create longer term effects to combat misinformation.

Finally, the path coefficients from the mediation analyses indicate that participants' recall of the event had a greater association with their attitude toward the relevant parties in the story at Time 2 than at Time 1. The mediation analysis also revealed that recall did not mediate the effect of debiasing messages on attitude until Time 2. In other words, the extent to which people's memory of an event determined their attitude toward the relevant parties in the event increased over time. These results suggest that one effective approach to shaping people's attitude toward an event in the long run is to identify strategies to enhance their comprehension of the correction message and their long-term memory of it.

Limitations and Conclusion

Despite its contributions, the current research has some limitations that should be addressed in future research. First, the study did not include a control condition where participants read the stimuli story without answering any questions. It is possible that the effects observed in Wave 2 of the study were due to a test effect rather than the persistent effect of the manipulations. This possibility should be checked in future research.

Moreover, in the present study, the corrective message was presented simultaneously with the misinformation and the debiasing information. Previous research discovered that delaying the exposure to correction messages led participants to treat the correction as additional information rather than contradictory information to the misinformation, given that people's memory of the misinformation deteriorates over time (Moore & Lampinen, 2016). Given these results, it is possible that, after being exposed to misinformation, a long delay before the exposure to corrective messages will weaken people's ability to integrate the corrective message and the original misinformation, thereby impeding the outcome of misinformation correction. Future research should therefore explore the time span between misinformation and its correction as a potential moderator or boundary condition for the message order effect discovered in the present study.

It is also important to note that the present study tested the order of misinformation and correction using scenarios in which negating the misinformation does not lead to an affirmative counterpart. It is unclear whether the same findings will replicate in cases in which an affirmative counterpart is readily available. Future research should seek to replicate this effect using other types of misinformation correction scenarios in which the comprehension of a correction does not require the coactivation of the negation tag and the negated information. It is also worthwhile for future research to consider negation comprehension mode as a potential moderator in combating the continued-influence effect of misinformation.

Moreover, several design decisions were made to maximize the internal validity of the design, potentially at the expense of external validity. First, the experimental manipulations took place in the vacuum of participants' preexisting attitude to minimize the random variances in the study and enhance the internal validity of the findings. However, it is possible that an individual who holds an attitude inconsistent with the correction message may change his or her attitude to the even more erroneous direction. Or, there may be a delayed effect of the correction message on individuals who initially hold an inconsistent attitude. It is hence worthwhile for future research to replicate the design using controversial issues to explore the boundary conditions of the findings. Second, the experimental stimuli featured fictitious stories. These stories were presented to the participants as plain text without any context. It is unclear whether the discovered effects would hold if the stories were presented as, for example, news stories published by certain news outlets on particular platforms. The credibility of the source and the medium may be meaningful moderators of the effects discovered in the present research, which should be explored in the future research.

To conclude, using a two-wave online experiment with samples collected from Amazon's MTurk, our study led to three major findings regarding misinformation correction. First, a correction is more effective when it is presented after rather than before the misinformation. This effect was shown to be robust, with a built-in replication in our design using two different versions of a story. Second, presenting a correction message together with a message that explains why the misinformation existed in the first place helps to enhance the memory of the correction message and, subsequently, the effectiveness of the correction message. Third, the effects of the experimental factors lasted for at least one week after the experiment even though people's memory about the stimuli diminished over time. The findings shed light on the design of misinformation correction strategies both in terms of its format and its content as well as provide guidance on news writing practices.

References

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. https://doi.org/10.1371/journal.pone.0118093

Blanc, N., Kendeou, P., van den Broek, P., & Brouillet, D. (2008). Updating situation models during reading of news reports: Evidence from empirical data and simulations. *Discourse Processes*, *45*, 103–121. https://doi.org/10.1080/01638530701792784

Blondé, J., & Girandola, F. (2016). Revealing the elusive effects of vividness: A meta-analysis of empirical evidence assessing the effect of vividness on persuasion. *Social Influence*, *11*(2), 111–129. https://doi.org/10.1080/15534510.2016.1157096

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 133–139). Washington, DC: American Psychological Association. https://doi.org/10.1037/14805-009

Crozier, W. E., & Strange, D. (2019). Correcting the misinformation effect. *Applied Cognitive Psychology*, *33*(4), 585–595. https://doi.org/10.1002/acp.3499

Devine, P. G., & Ostrom, T. M. (1985). Cognitive mediation of inconsistency discounting. *Journal of Personality and Social Psychology*, *49*(1), 5–21. https://doi.org/10.1037/0022-3514.49.1.5

Ecker, U. K. H., & Ang, L. C. (2019). Political attitudes and the processing of misinformation corrections. *Political Psychology*, *40*(2), 241–260. https://doi.org/10.1111/pops.12494

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. https://doi.org/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*(8), 1087–1100. https://doi.org/10.3758/MC.38.8.1087

Ecker, U. K. H., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2020). The effectiveness of short-format refutational fact-checks. *British Journal of Psychology*, *111*(1), 36–54. https://doi.org/10.1111/bjop.12383

Gordon, A., Brooks, J. C. W., Quadflieg, S., Ecker, U. K. H., & Lewandowsky, S. (2017). Exploring the neural substrates of misinformation processing. *Neuropsychologia*, *106*, 216–224. https://doi.org/10.1016/j.neuropsychologia.2017.10.003

Hameleers, M. (2020). Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the U.S. and Netherlands. *Information, Communication & Society*. Advance online publication. https://doi.org/10.1080/1369118X.2020.1764603

Hameleers, M., & van der Meer, T. G. L. A. (2020). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research*, *47*(2), 227–250. https://doi.org/10.1177/0093650218819671

Hardisty, D. J., Johnson, E. J., & Weber, E. U. (2010). A dirty word or a dirty world? Attribute framing, political affiliation, and query theory. *Psychological Science*, *21*(1), 86–92. https://doi.org/10.1177/0956797609355572

Hayes, A. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York, NY: Guildford Press.

Jang, J., Lee, E. J., & Shin, S. Y. (2019). What debunking of misinformation does and doesn't. *Cyberpsychology, Behavior, and Social Networking*, *22*(6), 423–427. https://doi.org/10.1089/cyber.2018.0608

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Kendeou, P., & O'Brien, E. J. (2014). The Knowledge Revision Components (KReC) framework: Processes and mechanisms. In *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 353–377). Cambridge, MA: MIT Press.

Lewanowsky, 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. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Maciuszek, J., & Polczyk, R. (2017). There was not, they did not: May negation cause the negated ideas to be remembered as existing? *PLoS ONE*, *12*(4), e0176452. https://doi.org/10.1371/journal.pone.0176452

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. https://doi.org/10.1080/10584609.2017.1334018

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs. "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40*(4), 433–449. https://doi.org/10.1016/j.jesp.2003.07.008

McGregor, I., & Holmes, J. G. (1999). How storytelling shapes memory and impressions of relationship events over time. *Journal of Personality and Social Psychology*, *76*(3), 403–419. https://doi.org/10.1037/0022-3514.76.3.403

Molina, M. D., Sundar, S. S., Le, T., & Lee, D. (2019). "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist*. Advance online publication. https://doi.org/10.1177/0002764219878224

Moore, K. N., & Lampinen, J. M. (2016). The use of recollection rejection in the misinformation paradigm: Recollection rejection of misinformation. *Applied Cognitive Psychology*, *30*(6), 992–1004. https://doi.org/10.1002/acp.3291

Rapp, D. N., & Braasch, J. L. G. (2014). *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. Cambridge, MA: MIT Press.

Reuters. (2019, April 2). *Factbox: "Fake News" laws around the world.* Retrieved from https://www.reuters.com/article/us-singapore-politics-fakenews-factbox/factbox-fake-news laws-around-the-world-idUSKCN1RE0XN

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? In B. Ross (Series Ed.), *The psychology of learning and motivation* (Vol. 41, pp. 265–292). Cambridge, MA: Academic Press. https://doi.org/10.1016/S0079-7421(02)80009-3

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

van den Broek, P., Risden, K., & Husebye-Hartmann, E. (1995). The role of readers' standards for coherence in the generation of inferences during reading. In R. Lorch & E. O'Brien (Eds.), *Sources of coherence in reading* (pp. 353–373). Hillsdale, NJ: Erlbaum.

van der Meer, T. G. L. A., & Jin, Y. (2020). Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health Communication*, *35*(5), 560–575. https://doi.org/10.1080/10410236.2019.1573295

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564

Weber, R., & Popova, L. (2012). Testing equivalence in communication research: Theory and application. *Communication Methods and Measures, 6*(3), 190–213. https://doi.org/10.1080/19312458.2012.703834

Appendix: Stimuli for the Negation-Driver Responsible-Debiasing Condition

[Messages 1–3 are filler messages.][2]

[4] *An earlier police statement pointed out that the accident happened because (a) the driver was distracted by a standing passenger on board. In trying to warn the passenger, the driver failed to notice a stop sign and crashed into the embankment.*

[Messages 5–13 are filler messages.]

[14] *Earlier investigation pointed to another cause of the accident, but it turned out to be untrue. Rather, the police found that the accident happened because (b) the driver did not signify properly before making a turn. A car from the next lane failed to respond in time and hit the van from behind.*

[15] *The driver/passenger gave false information to the police in an attempt to avoid taking responsibility for the accident, which led to the misleading conclusion in the earlier investigation.*

Notes. Italicized text reflects the experimental inductions. Messages (a) and (b) were switched in conditions where the passenger was the responsible party. Messages [4] and [14] were switched to in conditions where the corrective messages were presented first. Message [15] followed message [14] only in conditions that presented debiasing information.

[2] The full stimulus article is available at:
https://www.dropbox.com/s/y2jyh6k01vbxjtg/Full%20Appendix.docx?dl=0.