American Psychological Association

Technology, Mind & Society
2021 Conference Proceedings

ISBN:

http://dx.doi.org/10.1037/xxxxxxx

(In)effectiveness of Accumulated Correction on COVID-19 Misinformation

Haeseung Seo, Aiping Xiong, Sian Lee, Dongwon Lee
The Pennsylvania State University

An effective correction on COVID-19 misinformation is necessary for improving public health. To explore the effects of various methods to correct misinformation on social media, we examined the effects of accumulated corrections (e.g., one vs. two. vs. three) by two types of social-media users (e.g., individuals vs. health organizations) on COVID-19 fake news. We found that participants tended to reduce their perceived accuracy ratings and willingness to share misinformation with correction compared to a control condition. However, a significant effect of accumulated corrections was not observed. To understand the possible reasons behind the ineffectiveness, we did an exploratory analysis on expressional types of correcting comments and found that *the simpler the comment is, the more effective the correction is*. Our findings suggest making correcting comments "simple" in terms of COVID-19 fake news on social media.

Keywords: Misinformation, Fake news, COVID-19, Correction, Social media

Since the onset of the pandemic, much misinformation about COVID-19 has been generated and spread on social media. Fake news related to COVID-19 even made some people die from wrong treatments. To protect public health from such misinformation, we need to explore practical ways to "correct" fake news. Given users' leading roles in information-sharing on social media (Boyd et al., 2007), *user-initiated* correction can be an effective approach to encouraging users' active involvement in filtering out suspicious information,

In particular, Vraga and Bode (2017) found the effect of correction depending on the source and the number of corrective responses. Their study showed the impacts of correction from CDC and a user followed by CDC, respectively. However, their study was limited in using only one piece of fake news and correction from one health organization. They did not distinguish the types of correction contents when examining the number effect. In addition, they recruited undergraduate students and evaluated their misperceptions of the fake news before and after the correction(s) instead of directly examined their judgment on the fake news.

Method

The current study aims to address those limitations in Vraga and Bode (2017) by further exploring the effect of correction comments depending on the source and accumulation through an online human-subject experiment.

Participants. We recruited participants by posting Human Intelligent Tasks (HITs) on Amazon Mechanical Turk. We recruited only the workers who (1) were at least 18 years old; (2) were located in the U.S.; and (3) completed more than 100 HITs with a HIT approval rate of at least 95%. Qualtrics was used to program our online studies. Our study was approved by the institutional review board (IRB) office at the authors' institution.

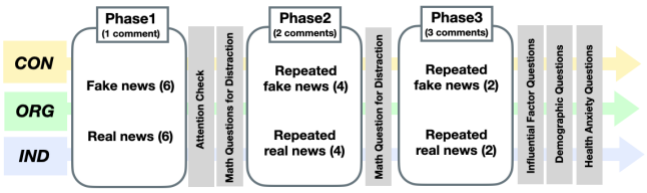
Materials. We selected twelve news articles about COVID-19 released from May to July 2020 from reputable fact-checking websites, i.e., *snopes.com* or *politifact.com*. Half of the articles were fake, and the other half were real. In addition, another piece of real news was selected for attention check. We created a simulated Twitter interface, where each piece of news was embedded within a tweet message. For each stimulus, a tweet message was shown above the COVID-19 news. The tweet message was a short sentence related to the news without any correcting message. The embedded news included an image, a news headline, and a snippet of the news content. Then, a comment from another user was presented under the news article. Each piece of fake news had a correcting comment(s) with a reference URL, while each piece of real news only had a comment complying with the news claim.

There were three conditions regarding correction for fake news: no correction (*CON*), correction by individuals (*IND*), and correction by health organizations (*ORG*). The same correcting message and reference URL were shown for both *ORG* and *IND*, while a comment not containing a correcting message or a reference URL was shown for *CON*. Regarding accumulated corrections, we created three types of correcting comments (Lewandowsky et al., 2012): 1) Simple, Brief Rebuttal (*R*): the simplest form to debunk

This article was published [to be completed by publisher].
This research was supported in part by the PSU SSRI seed grant and NSF awards #1742702, #1820609, and #1915801. The authors declared no potential conflict of interest with respect to research, authorship, and/or publication of this article.
Correspondence concerning this article should be addressed to Aiping Xiong, College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, 16802, the U.S. Email:axx29@psu.edu

fake news with a negating sentence; 2) Emphasis on Facts & Simple, Brief Rebuttal (*ER*): a sentence emphasizing the falsity of fake news, followed by a negating sentence; 3) Alternative Account (*A*): an explanation for trying to fill the gap left by retracting misinformation. We also varied the reference URLs from three health organizations (CDC, WHO, or NIH).

Figure1. ***A Flow Chart of the Experiment. CON, ORG, IND Refer to the Three Between-Subject Conditions***



Procedure. After informed consent, participants were randomly assigned to one of the three conditions (see Figure 1). In each condition, they evaluated six pieces of fake news and six pieces of real news at Phase 1. We presented four pieces of the fake news and four pieces of the real news from Phase 1 "again" at Phase 2. At Phase 3, participants viewed half of the fake news and half of the real news of Phase 2. To examine the effect of accumulated corrections, we appended a different type of correcting comment below the previous one to each piece of fake news across phases, e.g., *R* (Phase 1) →*ER* (Phase 2) →*A* (Phase 3). We implemented a semi-Latin square design to assign the comment of each type in a relatively balanced way.

To measure participant's acceptance of the "claim" of embedded news in each tweet, we first asked them to answer, "How accurate is the claim in the above news?" on a 7-point scale with "1" meaning "Very inaccurate" and "7" meaning "Very accurate." Then, they rated their willingness to share the news by answering, "Would you consider sharing this news online (for example, through Facebook or Twitter)?" using another 7-point scale with "1" meaning "Never" and "7" meaning "Always." Participants answered the same two questions for each piece of news in each phase. At the end of Phase 1, an extra piece of real news was included to exclude inattentive participants. Moreover, we inserted two simple math questions between phases to prevent maintenance of previous corrections from participants' working memory. After all phases, there was a post-session questionnaire about influential factors and demographic information.

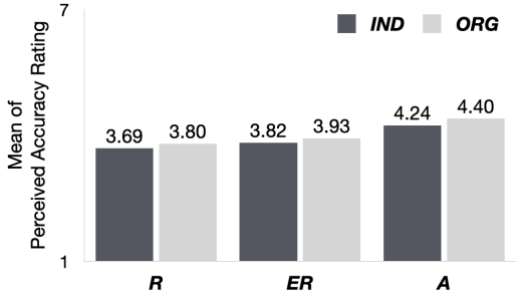
Results

We adopted 664 valid submissions for data analysis: 218 (*CON*), 211 (*IND*), and 235 (*ORG*). We paid $1.9 for participants who completed the task (equivalent to the hourly payment of $7.6). Perceived accuracy rating and willingness-to-share measure were entered into 3 (condition: *CON, IND, ORG*) × 2 (veracity: *Fake, Real*) × 2 (accumulation: *One, Two, Three*) mixed analysis of variances (ANOVAs) with a significance level of .05, respectively. Post-hoc tests with Bonferroni correction were performed. We report the effect size η_p^2 using SPSS. Mean values of two pieces of news were consistently used for checking the accumulation effect across the three phases.

Participants clearly distinguished real news (5.15) from fake news (4.02), $F(1,661) = 275.48$, $p < .001$, $\eta_p^2 = .294$. The interaction of news veracity × condition only approached significance, $F(2,661) = 2.54$, $p = .079$, $\eta_p^2 = .008$. Participants' willingness to share real news (4.44) was higher than that of fake news (3.84), $F(1,661) = 130.63$, $p < .001$, $\eta_p^2 = .165$. The interaction of news veracity × condition was significant, $F(2,661) = 3.47$, $p = .032$, $\eta_p^2 = .010$, but the follow-up pairwise comparisons did not show any significant difference. No effect involving accumulation was significant.

Next, to better understand the limited effect of correction by sources and accumulation, we did an exploratory analysis on the perceived accuracy ratings at Phase 1 by entering them into 3 (type: *R, ER, A*) × 2 (condition: *IND, ORG*) mixed ANOVA. There was a main effect of correction type, $F(2,444) = 54.94$, $p < .001$, $\eta_p^2 = .110$. Post-hoc analysis revealed that for both conditions, *R* and *ER* reduced participants' perceived accuracy ratings more than *A*, respectively (*ps* < .001, see Figure 2). The reduction of *R* showed a trend to be larger than that of *ER* ($p = .064$).

Figure2. ***Mean Values of Perceived Accuracy Ratings by Type × Condition for the Fake News.***



The post-session questionnaire results unearthed that the majority of *ORG* (32.7%) and the second majority of *IND* (28.6%) chose "other's comments" as the most influential factor(s) in evaluating the perceived accuracy ratings on a piece of fake news. Furthermore, as the most influential reason to choose the "others' comments," participants in *ORG* chose "who wrote the comment" the most (40.6%) and participants in *IND* chose "whether the comment included a reference URL" the most (37%), $X^2(4) = 39.37$, $p < .001$.

Discussion

We examined the effects of accumulated correction comments from different sources. We observed the limited effects of correction by source in participants' perceived accuracy rating but obtained that the participants counted on the reliability of correction in deciding their perceived accuracy rating. Our exploratory analysis revealed that the ineffectiveness might be due to the impact of the type of

correcting comments. Moreover, our results implied that a "simpler" comment (e.g., *R*) had the greater correction effect, suggesting the importance of linguistic simplicity in warning online (Harbach et al., 2013).

References

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*, 210-230.

Harbach, M., Fahl, S., Yakovleva, P., & Smith, M. (2013). Sorry, I don't get it: An analysis of warning message texts. In *International Conference on Financial Cryptography and Data Security* (pp. 94-111). Springer, Berlin, Heidelberg

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 Interes*t, *13*, 106-131.

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