The Harvard Kennedy School Misinformation Review[1]
June 2020, Volume 1, Issue 4

DOI: https://doi.org/10.37016/mr-2020-026
Website: misinforeview.hks.harvard.edu

Research Article

Do the right thing: Tone may not affect correction of misinformation on social media

An experiment conducted with 610 participants suggests that corrections to misinformation – pointing out information that is wrong or misleading and offering credible information in its place – on social media reduce misperceptions regardless of the correction's tone (uncivil, affirmational, or neutral). There is also an opportunity to correct secondary but related misperceptions (dealing with the same topic but with a different specific fact) when responding to misinformation on social media. Our findings emphasize that correction on social media could operate as part of a broader strategy to reduce beliefs in misinformation, and users should be encouraged to bring additional relevant information into the conversation, using whatever tone feels most comfortable for them.

Authors: Leticia Bode (1), Emily K. Vraga (2), Melissa Tully (3)
Affiliations: (1) Communication, Culture and Technology, Georgetown University, (2) Hubbard School of Journalism and Mass Communication, University of Minnesota, (3) School of Journalism and Mass Communication, University of Iowa
How to cite: Bode, L.; Vraga, E. K.; Tully, M. (2020). Do the right thing: Tone may not affect correction of misinformation on social media, *The Harvard Kennedy School (HKS) Misinformation Review*, Volume 1, Issue 4
Received: April 22nd, 2020 Accepted: June 2nd, 2020 Published: June 10th, 2020

Research questions

- How does the tone of a correction of misinformation on social media affect misperceptions?
- Can a social media comment preemptively correct a common myth that is related to but distinct from the original misinformation?
- How does the tone of a correction on social media affect perceptions of adjacent posts?

Essay summary

- In this experiment, 610 participants from Amazon's Mturk platform were asked about their beliefs about the safety and nutrition of raw milk. Participants were shown a simulated Twitter feed.
- Participants were shown a meme that contained misinformation about raw milk's nutrition. Some were also shown a correction, in which the tone was either neutral, uncivil, or affirming, but the facts remained the same.

[1] *A publication of the Shorenstein Center for Media, Politics, and Public Policy, at Harvard University, John F. Kennedy School of Government.*

- Those who saw a correction experienced reduced misperceptions (they had lower beliefs in the misinformation about raw milk) offering more evidence that observational correction – that is, watching someone else get corrected – is effective at reducing misperceptions. This effect was consistent whether they saw a neutral (factual-only), uncivil, or affirmational correction.
- Correcting a related misperception – dealing with the safety, rather than the nutritional value, of raw milk – reduced misperceptions on that issue.
- Those who saw the uncivil correction perceived it as less civil, and also thought the original post was less civil, suggesting a spillover incivility effect, in which perceptions of incivility seem to transfer to adjacent social media content.
- Interventions aimed at encouraging and facilitating user-to-user correction on social media should emphasize content over tone, and encourage users to correct in whatever tone feels most comfortable to them, as long as they provide factual information.

Implications

Research increasingly shows that correcting misinformation is effective at getting people to update their beliefs – when you give them new facts, they tend to reduce their beliefs in misinformation (Porter & Wood, 2019; Walter & Murphy, 2018). This works in a variety of contexts, including on social media, where misinformation often spreads quickly but can also effectively be corrected, including by other social media users (Bode & Vraga, 2018; Vraga & Bode, 2017; 2018). This is particularly promising, given that the visible and networked nature of social media allows for *observational correction* – users on social media can watch others be corrected (Vraga & Bode, 2017), increasing the impact of any given correction as it is seen by an entire network of users.

In this study, we test two open questions from this line of research. First, does the tone of a correction on social media – specifically, whether users adopt a neutral tone, in which the correction posts simply convey a fact, an uncivil tone, in which the correction posts insult the original poster, or an affirmative tone, in which correction posts show empathy and affirm the original poster, when responding to misinformation – increase or decrease its ability to reduce misperceptions (Chen, 2017; Ecker, Swire, & Lewandowsky, 2014)? And second, can users correct *related* topics of misinformation, not mentioned in the original post, while correcting a direct claim made by a social media user?

Perhaps our clearest takeaway is that tone does not influence how effective corrections are. In other words, a correction that focuses purely on offering a factual response to misinformation is equally successful in reducing misperceptions as one that is uncivilly attacking the intelligence of the person sharing the misinformation. For this reason, 'fixing' the problem of incivility on social media, while it may be admirable on its own, should not be expected to impact the problem of misinformation and misperceptions resulting from social media. Although uncivil replies do not make corrections less effective, the tone changes perceptions of the original tweet making it appear more uncivil. In other words, when people see an uncivil reply, they perceive it as such and also think the original post was more uncivil, suggesting a possible spillover effect (Anderson, Brossard, Scheufele, Xenos, & Ladwig, 2014). Although uncivil corrections are equally effective in reducing misperceptions, they may negatively influence people's perceptions of the broader dialogue, which may prove consequential for other outcomes beyond correction.

Likewise, an affirmative tone – which acknowledges and affirms that confusion or uncertainty on the issue is understandable – is neither more nor less effective in reducing misperceptions than a neutral or

uncivil tone. While an affirmative tone may make the correction more palatable for the person getting corrected or for the individual engaging in the correction, it may not affect how the broader social media community *witnessing* the correction interprets or accepts the corrective information. It is worth noting that although affirmation is not simply an abundance of civility, our participants did not see the affirmative correction as more civil than a neutral correction – both were perceived as civil. Future research should examine whether stronger affirmations are perceived as more empathetic and less threatening than other types of corrections, which we did not test directly, and if this leads to more success in reducing misperceptions.

These findings have several implications. First, while we do not encourage social media users to be rude, uncivil corrections that include the same factual information are not less effective at reducing misperceptions. Thus, if a social media user feels more comfortable correcting with a somewhat rude reply, they should not worry that doing so will make their reply less effective in reducing misperceptions among bystanders seeing the interaction. Likewise, affirmative corrections do not appear to make *audiences* seeing the interaction – who may hold misperceptions themselves – more receptive to the correction. If being empathetic and affirming makes offering a correction easier, users should feel empowered to do so. Altogether, this suggests that those engaging in correction on social media have the flexibility to adopt the tone they think most appropriate. Users might consider the target of the misinformation, community norms, or their own comfort when selecting what approach to take in responding to misinformation (Tandoc, Lim & Ling, 2020). For those creating interventions to increase corrective efforts on social media, like media literacy groups and professional fact checkers, greater attention should be given to the *content* of the message – including links and supportive information (Vraga & Bode, 2018) – as compared to the tone of that content, to maximize the volume and effectiveness of corrections. Journalists, public health authorities, or other experts can also employ these strategies – sharing credible and relevant information in direct response to user misinformation on social media.

Of course, there is potential for the tone of a correction to affect other outcomes beyond misperceptions, which we do not explicitly test. If an uncivil tone leads to disengagement with the issue or an affirmative tone makes it more likely users are willing to engage in correction of those spreading misinformation change their mind on the issue, this could change recommendations on appropriate tone to be used. Likewise, an uncivil tone may be more problematic (or an affirmative tone more successful) for more emotional, salient, politicized, or partisan issues (Bolsen & Druckman, 2018), which future research should test.

Opportunities should also be taken to preemptively target related misperceptions when correcting on a particular topic. Essentially this allows correction to be doubly effective – correcting the specific piece of misinformation espoused on social media, but also a piece of related misinformation at the same time. In this study, the misinformation post only discussed the nutritional value of unpasteurized milk. However, public health officials tend to be more concerned with the health risks of unpasteurized milk (CDC, n.d.). Correcting related misperceptions – in our case, about the health risks – can therefore encourage appropriate behaviors based on more complete information. This preemptive correction may be particularly effective when multiple topics of misinformation are prominent for a given topic, as is often the case. More specifically, this might mean that social media platforms pair multiple corrections, rather than just surfacing one most directly related to misinformation a user has posted. For example, Facebook currently shows users related articles with fact checks when they see a misinformation post that has been identified as such by third-party fact checkers (Lyons, 2017). They might consider adding a second related article, debunking a related myth, in order to maximize the corrective effects on users.

Findings

Finding 1: User corrections decrease people's belief in misinformation regarding the nutritional value of raw milk, and this is true for all corrections regardless of tone

Participants who saw any of the corrections expressed lower levels of misperceptions that raw milk is more nutritious than pasteurized milk as compared to people who saw the misinformation without any correction. Moreover, there are no differences in misperceptions about raw milk's nutritional value among people who saw a neutral (factual-only) correction, an uncivil correction, or an affirmative correction.

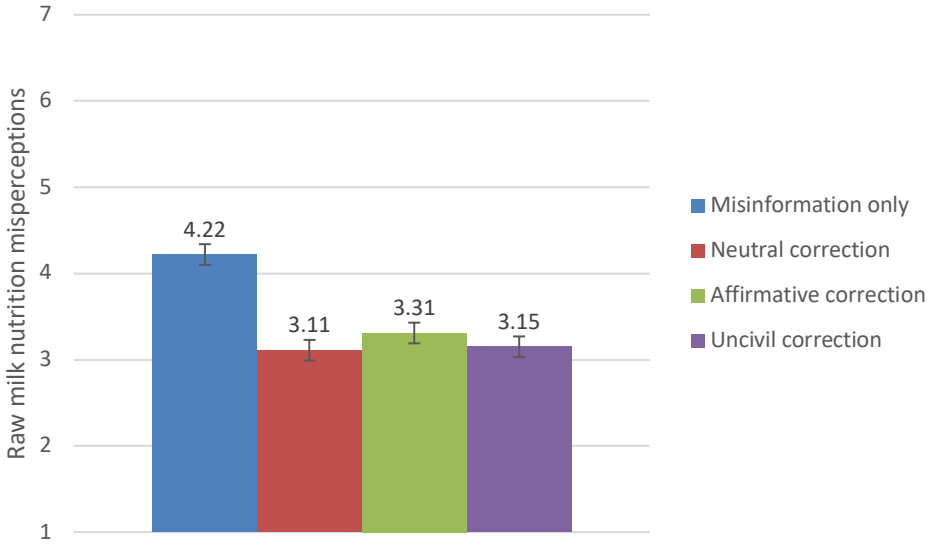

Figure 1. Misperceptions about the Nutritional Value of Raw Milk by Experimental Condition. *Note that higher numbers reflect less accurate attitudes towards the nutritional value of raw milk. Each color shows the average for the group. Error bars represent a 95% confidence interval for the results.*

Finding 2: Corrections also reduce people's beliefs in a related piece of misinformation (that raw milk is safe to drink), and this is true for all corrections

The original misinformation post did not mention the safety of raw milk, nor did the first correction. Instead, this idea only emerged in the second correction, which not only debunked the original misinformation but also added that pasteurized milk is safer to drink than raw milk. Our analyses confirm that exposure to *any* of the corrections (regardless of tone) reduced misperceptions regarding raw milk's safety as compared to people who only viewed misinformation. Moreover, there are no differences in terms of raw milk misperceptions among the three corrections themselves; misperceptions in terms of raw milk's safety are equally low regardless of the tone of the correction.

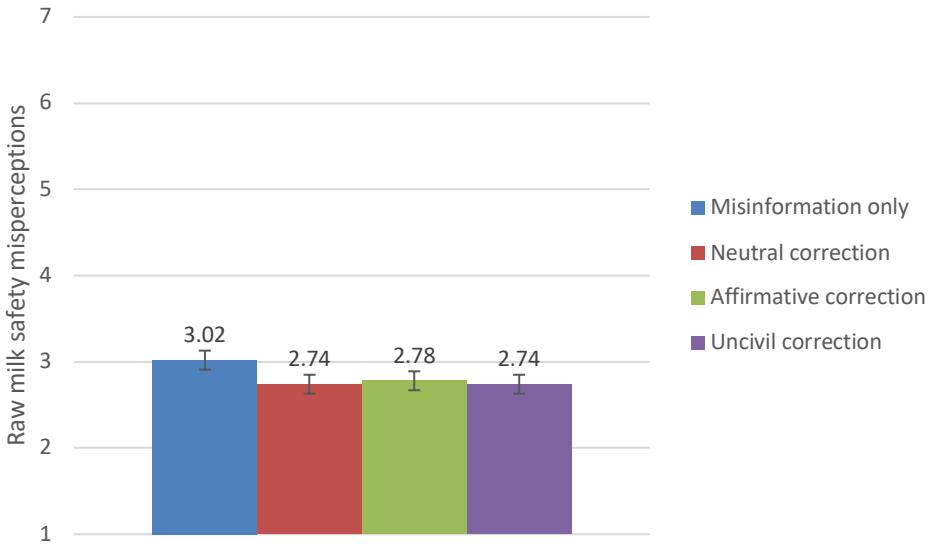

Figure 2. Misperceptions about the Safety of Raw Milk by Experimental Condition. *Note that higher numbers reflect less accurate attitudes towards the safety of raw milk. Each color shows the average for the group. Error bars represent a 95% confidence interval for the results.*

Finding 3. Participants thought the uncivil corrections were uncivil

When rating the civility of the corrections, the uncivil corrections were seen as more uncivil than either the affirmative or neutral corrections, which were seen as equally civil. However, even the uncivil corrections were only seen as moderately uncivil, receiving a score of 4.08 out of a possible 7 in terms of the civility of the corrections. Notably, the affirmative corrections were not seen as more civil than the neutral corrections: both were seen as more civil than not and as more civil than the uncivil correction. However, this measure (perceived civility) is not designed to capture affirmation which is about empathy and threat reduction, so we are cautious in our interpretation.

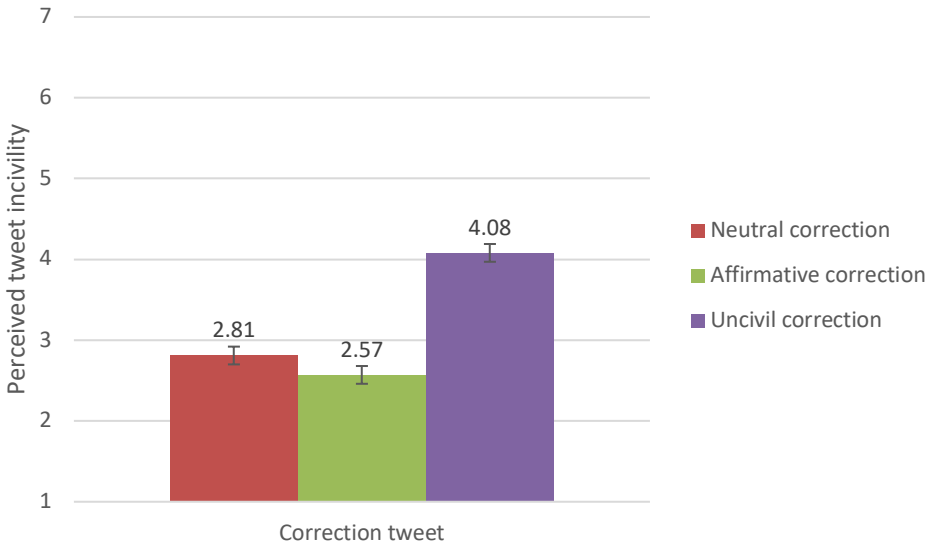

Figure 3. Perceptions of Correction Incivility by Experimental Condition. *Note that higher numbers reflect greater perceptions of incivility. Each color shows the average for the group. Error bars represent a 95% confidence interval for the results.*

Finding 4. Uncivil corrections led people to believe the original tweet was also more uncivil, even though it was not

The original misinformation tweet, which was consistent across all the content seen by participants, was seen as more uncivil when the corrections were uncivil than when that misinformation tweet appeared absent any corrections. Meanwhile, the neutral and affirmative corrections fell between these extremes in terms of perceptions of their civility. In general, however, people did not perceive any of the tweets or replies to be particularly uncivil.

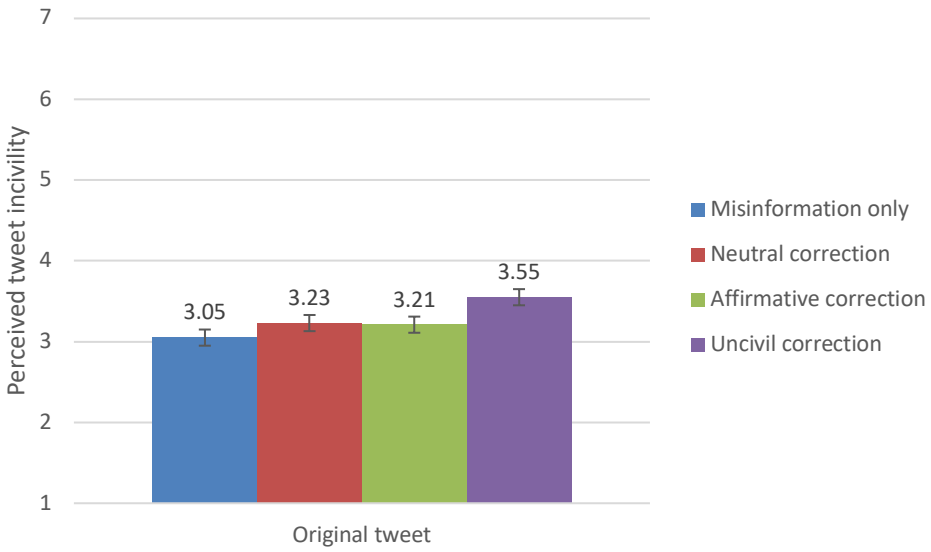

Figure 4. Perceptions of Original Tweet Incivility by Experimental Condition. *Note that higher numbers reflect greater perceptions of incivility. Each color shows the average for the group. Error bars represent a 95% confidence interval for the results.*

Methods

An experiment is the best way to test our hypotheses and research questions because it lets us isolate the specific effects of the tone of correction on misperceptions of those seeing the interaction. Participants were recruited from Amazon's Mechanical Turk in September of 2018. Participants were an average of 36 years old, 55% male, and largely educated (52% had at least a Bachelor's degree).

After completing a short pre-test questionnaire, participants were randomly assigned to view one of four simulated Twitter feeds (*N*=610).[2] In each condition (that is, each group into which participants were randomly assigned), participants were instructed to read a page of Twitter posts as if they were viewing their own feed, which we said were taken from someone's feed. The simulated feed contained six Twitter posts, including one manipulated post and five posts validated as politically neutral and plausible social media posts (Authors, 2016). Participants were required to spend 15 seconds on the page of posts before they could continue with the survey. After answering a series of questions regarding their experience with the feed, their evaluation of the posts, and their attitudes towards the target issue (raw milk), participants

[2] We examine four of the 10 total conditions we fielded in this article, excluding a pure control condition, in which participants were not exposed to any misinformation or correction, as well as 5 conditions that included a second manipulation focused on news literacy. None of the conditions tested in this article included news literacy interventions.

were thanked for their participation and debriefed. Debriefing included a statement that pasteurized milk is equally nutritious and safer to consume than raw (unpasteurized) milk and provided a link to more information from the CDC. Participants were paid $1.10 for participating in the survey.

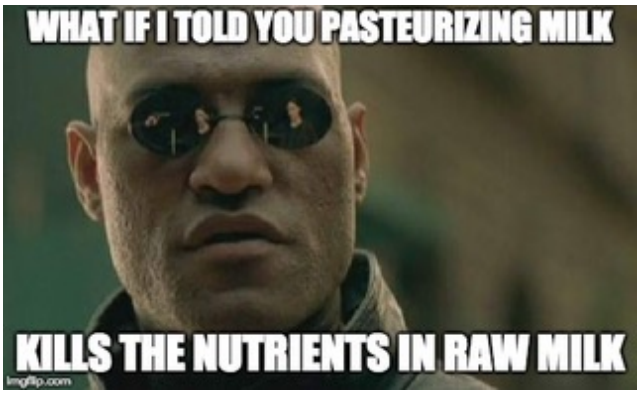

Figure 5: The Misinformation Meme that all Respondents Viewed

The third post on each feed contained the experimental manipulation. In all conditions, the same user posted a meme (an image macro featuring Morpheus from the Matrix) claiming that pasteurizing milk "kills the nutrients in raw milk" (see Figure 5). We selected this myth for several reasons. First, it is increasingly relevant, as raw milk is gaining attention and policy protections around the U.S. in recent years (Rahn, Gollust, & Tang, 2017). Second, the specific myth related to relative nutrition of raw and pasteurized milk is prominently debunked – the CDC and the FDA both have corrected this myth in their online materials (e.g., CDC, n.d.; FDA, n.d.). While overall consumption levels of raw milk (1.8% in the last seven days among respondents in the Midwest, Rahn, et al., 2017) and raw cheese (6.9% in the last seven days among respondents in the Midwest, Rahn, et al., 2017) are quite low, such consumption contributes to an outsized number of food-borne illness in the U.S. (60% of those reported in relation to dairy products were connected to unpasteurized milk according to Langer, et al., 2012). There is also at least moderate support for loosening restrictions on selling unpasteurized milk (52% supported loosening restrictions in a survey of Midwesterners, Rahn, et al., 2017).

In the misinformation-only condition, there were no responses to this post. In all correction conditions, two users responded to the post to debunk the misinformation, while providing links[3] to expert sources (e.g. the CDC and FDA), in keeping with best practices (e.g., Vraga & Bode, 2018). The first response included a link to the CDC and directly responded to the misinformation that pasteurization kills nutrients. The second response reinforced the equal nutritional value of raw and pasteurized milk and added that pasteurization keeps people from getting sick from bacteria in milk, providing a link to the FDA. This second claim is what we use to test whether 'related misperceptions' of raw milk safety are updated.

[3] Note that because they are simulated feeds, participants did not have the opportunity to click the links. No additional factual content is therefore associated with the presence of the links.

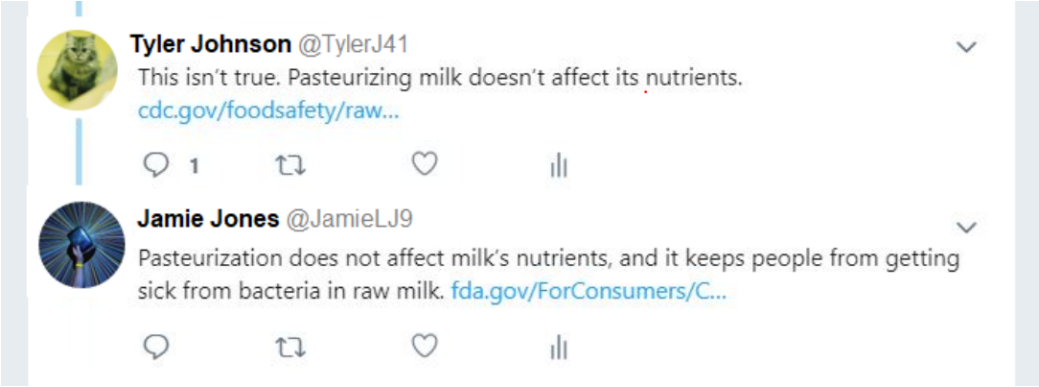

Figure 6: Neutral Correction (factual-only)

In the "neutral correction" condition, the responses provide a factual response and mimic the tone of earlier designs on observational correction (e.g., Bode & Vraga, 2018; Vraga & Bode, 2017; 2018). The specific text of the first reply was "This isn't true. Pasteurizing milk doesn't affect its nutrients. cdc.gov/foodsafety/raw…," and the second reply said: "Pasteurization does not affect milk's nutrients, and it keeps people from getting sick from bacteria in raw milk. fda.gov/ForConsumers/C." (Figure 6). In the "affirmative correction," the responses attempt to validate the participants' worldview and concerns, expressing understanding of possible "confusion" on the issue, as suggested by Lewandowsky et al. (2012). The text of the first affirmative reply said "I know it can be super confusing, but this isn't true. Pasteurization doesn't affect the nutrients in milk at all. cdc.gov/foodsafety/raw…" and the second reply said "This is such a scary thought, but I just learned that pasteurization does not affect milk's nutrients, and it keeps people from getting sick from bacteria in raw milk! fda.gov/ForConsumers/C." Finally, the "uncivil correction," includes insults and name-calling (Chen, 2017) towards the original poster in the correction, telling the poster not to be "stupid" and calling them an "idiot". The first uncivil reply said "Oh come on, don't be stupid. Everyone knows that pasteurizing milk doesn't affect its nutrients. cdc.gov/foodsafety/raw…" and the second reply said: "I can't believe how dumb this is. Pasteurization does not affect milk's nutrients, and it keeps people—even idiots like you—from getting sick from bacteria in raw milk. fda.gov/ForConsumers/C."Although this language could be considered a mild form of incivility (Chen, 2017), especially given the kind of language that circulates on Twitter (Oz et al., 2018; Phillips, 2015), it clearly attacks the Twitter user directly and is used as a means of undercutting their claim about raw milk (which in fact is incorrect).

To test the effects on misperceptions, we perform a series one-way ANOVAs, first comparing exposure to any correction (the three corrections combined) versus misinformation-only; then comparing among the three correction conditions. For the effects on civility perceptions, we compare between all four (original tweet civility) or three (correction civility) conditions using an omnibus test. For significant results, we use pairwise comparisons with a Bonferroni correction to adjust for additive error to examine differences between the conditions. Figures include the estimated marginal means for all relevant conditions to facilitate comparison. See the supplemental methods addendum for more information.

Bibliography

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "nasty effect:" Online incivility and risk perceptions of emerging technologies. Journal of Computer-Mediated Communication, 19(3), 373-387.

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health

misinformation on social media. Health communication, 33(9), 1131-1140.

Bolsen, T., & Druckman, J. N. (2018). Do partisanship and politicization undermine the impact of a scientific consensus message about climate change?. Group Processes & Intergroup Relations, 21(3), 389-402.

CDC. (n.d.). 5 Raw Milk Myths Busted! https://www.cdc.gov/foodsafety/rawmilk/milk-myths.html

Chen, G. M. (2017). Online incivility and public debate: Nasty talk. Cham, Switzerland: Palgrave Macmillan.

Ecker, U. K., Swire, B., & Lewandowsky, S. (2014). Correcting misinformation—A challenge for education and cognitive science. *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*, 13-38.

FDA. (n.d.). The dangers of raw milk: Unpasteurized milk can pose a serious health risk. https://www.fda.gov/food/buy-store-serve-safe-food/dangers-raw-milk-unpasteurized-milk-can-pose-serious-health-risk

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.

Lyons, T. (2017). Replacing Disputed Flags With Related Articles. Retrieved from: https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. New Media & Society, 20, 3400–3419. doi: 10.1177/1461444817749516

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture.* Boston, MA: MIT Press.

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

Rahn, W. M., Gollust, S. E., & Tang, X. (2017). Framing food policy: the case of raw milk. *Policy Studies Journal, 45*(2), 359-383.

Tandoc Jr., E., Lim, D., & Ling, R. (2020). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism, 21*(3), 381-398. doi: 10.1177/1464884919868325

Vraga, E. K., & Bode, L. (2017). Using Expert Sources to Correct Health Misinformation in Social Media. *Science Communication, 39*(5), 621-645. doi: 10.1177/1075547017731776

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

Funding
This project was funded by a Page and Johnson Legacy Scholars Grant, #2018FN004, from Pennsylvania State University.

Competing interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethics
The research protocol was approved by the institutional review board at George Mason University. Human subjects gave informed consent before participating and were debriefed at the end of the study.

Data Availability
All materials needed to replicate this study are available via the Harvard Dataverse https://doi.org/10.7910/DVN/IJKIN3.

Appendix: Detailed Analysis

Finding 1

The effect of correction on misperceptions regarding raw milk's nutritional value is tested using a one-way ANOVA comparing the three correction conditions combined versus the misinformation-only condition. The results from this ANOVA are significant, F(1,608)=52.97, $p<.001$, *partial* $\eta^2=.080$. The post-hoc comparisons confirmed that exposure to *any* correction produced lower misperceptions regarding the nutritional value of raw milk ($M=3.19$, *S.E.=*.07) than in the misinformation-only condition ($M=4.22$, *S.E.=*.12, $p<.001$). In addition, misperceptions among the three correction conditions are equivalent, as confirmed by a one-way ANOVA comparing the three correction conditions separately F(2,455)=.71, $p=.49$, *partial* $\eta^2=.003$. Likewise, no significant differences emerge between the three types of corrections when examining the pairwise comparisons using Bonferroni corrections.

Our outcome measure of raw milk misperceptions used two items: (1) raw milk is more nutritious than pasteurized milk, and (2) the pasteurization process kills key nutrients in milk, each measured on a seven-point scale from "strongly disagree" to "strongly agree." These items were averaged into an index, with a higher number indicating more misperceptions on raw milk nutrition ($r=.79$, $p<.001$, $M=3.45$, *S.D.=*1.58).

Finding 2

The effects of correction on misperceptions about raw milk safety were again tested using a one-way ANOVA. Again, exposure to *any* correction ($M=2.75$, *S.E.=*.07) reduces misperceptions regarding raw milk safety, F(2, 608)=4.18, $p=.04$, *partial* $\eta^2=.007$ as compared to the misinformation-only condition ($M=3.02$, *S.E.=*.11). Similarly, a one-way ANOVA comparing the three correction conditions showed no significant difference among those conditions F(2,455)=.03, $p=.97$, *partial* $\eta^2=.000$.

Our outcome measure of raw milk safety used two items: (1) it is safer to drink pasteurized milk than raw milk, and (2) drinking raw milk increases your risk of getting a foodborne disease, each measured on a seven-point scale from "strongly disagree" to "strongly agree." These items were averaged into an index and reversed so that a higher number indicating more misperceptions on raw milk nutrition ($r=.60$, $p<.001$, $M=2.82$, *S.D.=*1.40).

Finding 3

A one-way ANOVA compares the three correction conditions in terms of the perceived incivility of the correction tweet and finds significant differences among them, F(2,455)=55.38, $p<.001$, *partial* $\eta^2=.196$. Post-hoc Bonferroni comparisons confirm that the uncivil tweet is seen as significantly more uncivil ($p<.001$) than either the factual or affirmative correction, which were seen as equally civil ($p=.36$).

Our measure of perceived incivility averaged two items, which asked participants to rate whether the tweet was civil/uncivil and respectful/disrespectful on seven-point semantic differentials. These items were averaged to form an index, with a higher score indicating greater perceived incivility ($r=.77$, $p<.001$, $M=3.15$, *S.D.=*1.49).

Finding 4

A one-way ANOVA comparing the four experimental conditions is significant, which does suggest differences among the conditions F(3, 606)=4.25, $p=.01$, *partial* $\eta^2=.021$. A follow-up test of pairwise

comparisons using the Bonferroni adjustment suggests that only the comparison between the misinformation-only and uncivil conditions is significant ($p<.01$), with the other two conditions falling between these extremes.

Our measure of perceived incivility averaged two items, which asked participants to rate whether the tweet was civil/uncivil and respectful/disrespectful on seven-point semantic differentials. These items were averaged to form an index, with a higher score indicating greater perceived incivility ($r=.62$, $p<.001$, $M=3.26$, $S.D.=1.28$).