Harvard Kennedy School Misinformation Review[1]
July 2025, Volume 6, Issue 3

DOI: https://doi.org/10.37016/mr-2020-179
Website: https://misinforeview.hks.harvard.edu

Research Note

The small effects of short user corrections on misinformation in Brazil, India, and the United Kingdom

How effective are user corrections in combatting misinformation on social media, and does adding a link to a fact check improve their effectiveness? We conducted a pre-registered online experiment on representative samples of the online population in Brazil, India, and the United Kingdom ($N_{participants}$ = 3,000, $N_{observations}$ = 24,000). We found that in India and Brazil, short user corrections slightly, but often not significantly, reduced belief in misinformation and participants' willingness to share it. In the United Kingdom, these effects were even smaller and not significant. We found little evidence that fact-check links made user corrections more effective. Overall, our results suggest that short user corrections have small effects and that adding a fact-check link is unlikely to make user corrections much more effective.

Authors: Sacha Altay (1), Simge Andı (2), Sumitra Badrinathan (3), Camila Mont'Alverne (4), Benjamin Toff (5), Rasmus Kleis Nielsen (6), Richard Fletcher (7)
Affiliations: (1) Department of Political Science, University of Zürich, Switzerland, (2) Department of Politics, University of Exeter, UK, (3) School of International Service, American University, USA, (4) Faculty of Humanities & Social Sciences, University of Strathclyde, UK, (5) Hubbard School of Journalism & Mass Communication, University of Minnesota, USA, (6) Department of Communication, University of Copenhagen, Denmark (7) Reuters Institute for the Study of Journalism, University of Oxford, UK.

How to cite: Altay, S., Andı, S., Badrinathan, S., Mont'Alverne, C., Toff, B., Kleis Nielsen, R., & Fletcher, R. (2025). The small effects of short user corrections on misinformation in Brazil, India, and the United Kingdom. *Harvard Kennedy School (HKS) Misinformation Review*, 6(3).
Received: March 27th, 2025. Accepted: July 8th, 2025. Published: July 23rd, 2025.

Research questions

- How do user corrections influence the perceived accuracy of social media posts containing false COVID-19 information, and how do they influence participants' willingness to share them?
- Are user corrections more effective when they contain links to news organisations' fact checks?
- Do user corrections have effects beyond corrected posts?

Research note summary

- Our experimental design randomly assigned respondents to one of three conditions. Participants rated nine social media posts about COVID-19 (three true, six false). In the control condition, the posts had no comment. In the correction condition, four false posts included a short user

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

comment debunking them. In the correction with fact-check condition, the user comment included a link to a fact check.

- User corrections had small, and often non-significant, effects on the perceived accuracy of false posts and participants' willingness to share them. The effects were largest in India and Brazil, while they were smallest in the United Kingdom.
- In India and Brazil, the effect sizes of corrections ranged from 0.07 to 0.17 on the 4-point scale. In the United Kingdom, they ranged from 0.01 to 0.07 on the 4-point scale.
- The difference between corrections without a link and corrections with a link was not statistically significant.
- The corrections had no spillover effects on uncorrected true and false posts.

Implications

Professional corrections performed by fact checkers are effective at correcting misperceptions and belief in false claims (Porter & Wood, 2021, 2024). However, the scale and speed of misinformation production often outpace fact-checking efforts, resulting in corrections that arrive only after much of the damage has been done. Additionally, fact checks face significant dissemination problems. Few people voluntarily seek them out (Graham & Porter, 2025; Guess et al., 2020; Porter & Wood, 2024), and some major social media platforms have recently scaled back systems that automatically display fact checks alongside questionable posts (Kaplan, 2025). Moreover, in contexts where the main modes of information sharing are encrypted chat apps, professional fact checks cannot be issued at scale within these closed networks, not to mention the poor state of content moderation in non-English languages (Okong'o, 2025). In these contexts, the burden of providing fact checks increasingly falls on platform users themselves.

Past work has shown that user corrections (i.e., users correcting misinformation by refuting it on social media) can be effective at reducing misperceptions (Bode et al., 2024; Bode & Vraga, 2018; Yang et al., 2022). Yet, little is known about the effects of user corrections in Global South countries (for exceptions, see: Badrinathan & Chauchard, 2024; Blair et al., 2024)—although expert corrections have been shown to be effective here (Porter & Wood, 2021).

Our findings suggest that short user corrections have only small effects on the perceived accuracy of social media posts containing false COVID-19 information and participants' willingness to share it. In the United Kingdom, correction effects ranged from 0.01 to 0.07 on the 4-point scale, corresponding to reductions in belief and sharing of 1.2% to 8.4%. In Brazil, correction effects ranged from 0.07 to 0.11 on the 4-point scale, corresponding to reductions in belief and sharing of 5.7% to 11.2%. In India, correction effects ranged from 0.07 to 0.17 on the 4-point scale, corresponding to reductions in belief and sharing of 5.0% to 11.6%. The effects of corrections were particularly small in the United Kingdom, potentially because belief in COVID-19 misinformation was so low that corrections had almost no scope to be effective: Before being exposed to corrections, participants did not believe the COVID-19 misinformation and were not willing to share it online.

We also tested whether adding a link to a fact check from a news organization would strengthen the effect of corrections on the perceived accuracy of misinformation and participants' willingness to share it. The links may give credence to the correction, either by signaling that it is backed up by reliable sources or that there is evidence supporting the correction. Prior work suggests that some corrections are more effective than others. For instance, between-study evidence from a meta-analysis shows that corrections performed by experts are more effective than those performed by non-experts (Walter et al., 2020). Yet, within-study evidence, in which the source of the corrector is experimentally manipulated (e.g., in one condition the post is attributed to an expert while in the other it is not), tends to show that the content of corrections matters more than their source. For example, corrections performed by the World Health

Organization or anonymous Facebook users show similar effects (Vraga & Bode, 2021). In general, messages are more persuasive when they both come from sources people trust and when they are backed up by evidence (Mercier, 2020; Petty & Cacioppo, 1986).

However, we found that links to fact checks are unlikely to make user corrections more effective. As shown in Figure 3, the links to fact checks in user corrections were not merely hyperlinks. When we conducted the study in 2021, Facebook previewed these links, prominently displaying the title of the fact check and its source. This presentation potentially provided complementary information about why the information was false and clearly signaled that a reputable news outlet had refuted it and that there is evidence supporting the correction. Thus, the absence of clear added benefits of fact-check links can hardly be attributed to a lack of visibility or usefulness: They were clearly visible and contained relevant information.

In line with past work (Bode & Vraga, 2018; Coppock, 2023; Martel & Rand, 2024), in Appendix D, we show that the corrections were not more or less effective depending on the tendency to believe in conspiracy theories, trust in social media, or trust in the news. Moreover, while previous research (Pennycook et al., 2020) has shown that fact-checking warnings can have spillover effects on uncorrected posts (e.g., by increasing the perceived accuracy of false posts or decreasing the perceived accuracy of true posts), we found no evidence of spillover effects. In Appendix B, we show that the main conclusion of the article holds when excluding participants who failed the pre-treatment attention check—that is, the effects of user corrections are small, and adding a link to a fact check is unlikely to make them more effective.

The main limitation is that user corrections were short, and participants could not actually click on the links to the fact checks in the experiment. Longer, more detailed corrections and clickable links may have yielded stronger effects. Another limitation is that, like many interventions against misinformation, our treatments are bundled (Guess et al., 2024), meaning that the corrections with and without a link differ in many ways, and our experimental design does not allow us to isolate which specific feature is responsible for any observed effects. For example, comments with a link may not necessarily draw attention to a reputable source but simply make the correction more visible. The differences between our treatment conditions mirror actual platform design: as of June 2025, Facebook continues to display comments with and without links in the same manner as our experimental treatments. And many social media and messaging platforms, like LinkedIn or WhatsApp, also offer a similar link preview with a title and an image. Given the applied focus of our research, we prioritized an intervention that mirrors real-world platform design. Finally, our measures of accuracy and sharing may not reflect people's actual behaviors on social media. For instance, the mere fact of asking participants to rate the accuracy of a post shifts their attention to accuracy, which is unlikely to be top of mind for people when scrolling through social media. Moreover, it has been shown that prompting participants to think about accuracy increases their sharing discernment (Pennycook & Rand, 2021). It is also far from certain that changes in belief induced by corrections result in changed attitudes or behaviors (Porter & Wood, 2024). Regarding sharing, it is not clear whether self-reported measures of sharing are representative of people's actual sharing behaviors, given that most social media users are "lurkers" who avoid sharing news or information about politics and social issues (McClain, 2021).

A key implication of our work is that user corrections are no panacea and that efforts to fight misinformation cannot rest entirely on the shoulders of social media users. Effective interventions against misinformation require a combination of strategies as well as reaching and targeting vulnerable populations (Bak-Coleman et al., 2022; Brashier, 2024; Budak et al., 2024). Social media users and ordinary citizens can meaningfully contribute, but institutional and platform-level interventions are likely to be much more impactful. For instance, while it is important to find ways to motivate users to perform corrections online, it may be more important to change the affordances of social media to make those corrections more prominent and impactful. Here, it may be useful to distinguish between organically

occurring user corrections (like the corrections in this study) and institutionalized forms of user corrections that are integrated into platforms in ways that affect display decisions. And while the former likely have small effects, the latter may be more impactful. For example, initiatives along the lines of Community Notes (formerly Bird Watch) could, in theory, offer a promising model for users to write corrections while leveraging collective intelligence to filter the highest quality corrections for readers (Drolsbach et al., 2024; Martel et al., 2024; Renault et al., 2024). Such initiatives could display longer corrections and more prominently than organically occurring user corrections, which have shown to be effective at reducing the spread of misinformation under the right conditions (i.e., a politically balanced crowd; Drolsbach et al., 2024; Martel et al., 2024; Renault et al., 2024). While these efforts are not a direct substitute for professional fact-checking collaborations, they can complement institutional measures.

Findings

Finding 1: User corrections had small and inconsistent effects on perceived accuracy of false information.

We first tested whether user corrections decreased the perceived accuracy of COVID-19 misinformation relative to no corrections (see Figure 1). We report the effect of corrections on the 4-point scale (b) and the percentage change relative to the baseline compared to the control condition (Δ).

In the United Kingdom, corrections with a link (b = -0.07, p = .16, Δ = -8.2%) and without a link (b = -0.01, p = .84, Δ = -1.2%) had no statistically significant effects on belief in misinformation. In Brazil, corrections with a link (b = -0.10, p = .054, Δ = -6.9%) and without a link (b = -0.08, p = .11, Δ = -5.7%) had no statistically significant effects on belief in misinformation. In India, corrections with a link significantly reduced belief in COVID-19 misinformation (b = -0.16, p = .016, Δ = -10.8%), while corrections without a link did not significantly reduce it (b = -0.08, p = .26, Δ = -5.0%).

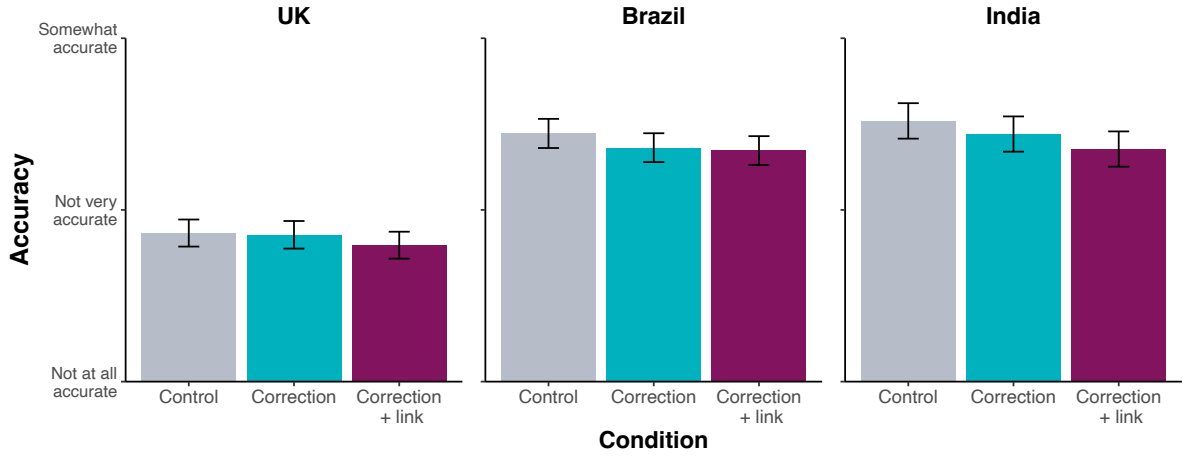

Figure 1. Bar plots representing the average accuracy ratings of false claims about COVID-19 in the Control condition (grey), Correction condition (blue), and Correction with a link to a fact check (purple). *The error bars represent the 95% confidence intervals. The full accuracy scale also includes a fourth point, "very accurate."*

Finding 2: User corrections had small and inconsistent effects on participants' willingness to share false information.

We tested whether user corrections decreased participants' willingness to share the posts containing false COVID-19 information relative to no corrections. In the United Kingdom, corrections with a link (b = -0.04, p = .46, Δ = -8.4%) and without a link (b = -0.01, p = .88, Δ = -1.8%) had no statistically significant effects on sharing intentions. In Brazil, corrections with a link (b = -0.07, p = .23, Δ = -7.4%) and without a link (b = -0.11, p = .051, Δ = -11.2%) had no statistically significant effects on sharing intentions. In India, corrections with a link (b = -0.17, p = .021, Δ = -11.6%) reduced participants' willingness to share COVID-19 misinformation, while corrections without a link did not significantly reduce it (b = -0.07, p = .32, Δ = -5.7%). Note that these differences between countries are not statistically significant, even when merging accuracy ratings and sharing intentions.

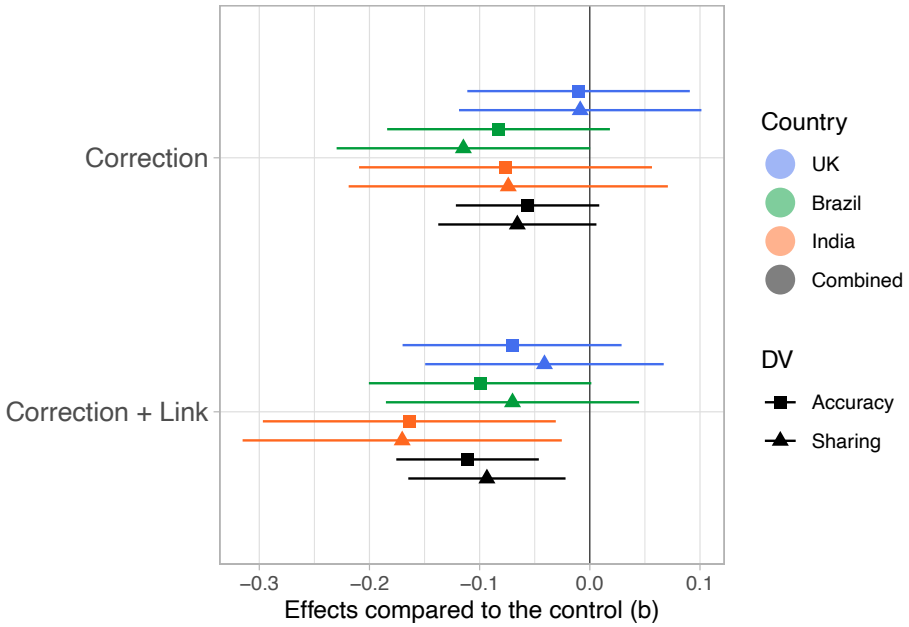

Figure 2. Effects of the corrections without a link and corrections with a link compared to the control condition (no correction). *The estimates (b) represent the treatment effects on the 4-point scale. The error bars represent the 95% confidence intervals.*

Finding 3: Corrections with a link to a fact check were not significantly more effective than corrections without it.

We tested whether corrections with a link to a fact check are more effective than corrections with no link. Using the combined data across countries, we did not find any evidence that corrections with a link to a fact check were significantly more effective than corrections without a link to a fact check at reducing belief in COVID-19 misinformation (b = -0.05, p = .10, Δ = -4.9%) and participants' willingness to share COVID-19 misinformation (b = -0.03, p = .45, Δ = -3.2%). The same is true in each individual country.

Finding 4: Corrections do not have spillover effects on uncorrected (true or false) posts.

We also investigated whether correcting some false posts, but not others, increases the perceived accuracy of the uncorrected false posts. Across countries, corrections did not have statistically significant

effects on the accuracy ratings of uncorrected false posts ($b_{no\ link}$ = 0.04, p = .29, Δ = 3.8%; b_{link} = -0.06, p = .14, Δ = -5.3%) or on participants' willingness to share uncorrected false posts ($b_{no\ link}$ = 0.06, 6.3%, p = .18; b_{link} = - 0.06, p = .13, Δ = -7.0%).

Second, we examined whether exposure to corrected false posts increases the perceived accuracy of the true posts. Across countries, corrections did not have statistically significant effects on the accuracy ratings of true posts ($b_{no\ link}$ = 0.001, p = .94, Δ = 0.1%; b_{link} = 0.04, p = .16, Δ = 1.9%) or on participants' willingness to share true posts ($b_{no\ link}$ = -0.01, p = .79, Δ = -0.7%; b_{link} = -0.02, p = .67, Δ = -1.2%). In Appendix C, we explore the determinants of belief in false COVID-19 and participants' willingness to share it.

Methods

Participants

We accessed Kantar Media's online survey panels to recruit 1000 participants in the United Kingdom (52% women, $Mdn_{age\ group}$ = 45–55, $Mdn_{education}$ = post-secondary), Brazil (55% women, $Mdn_{age\ group}$ = 35–44, $Mdn_{education}$ = short-cycle tertiary education, i.e., about two years after high school) and India (46% women, $Mdn_{age\ group}$ = 25–34, $Mdn_{education}$ = short-cycle tertiary education). Data collection took place in the United Kingdom and India, March 12–17, 2021, while in Brazil it was conducted March 12–24, 2021. The participants were distributed across the Control, Correction, and Correction with Link conditions as follows: (United Kingdom) 337, 321, 342; (Brazil) 326, 336, 338; (India) 330, 335, 335. Per country, this sample size allowed us to reliably detect effect sizes as small as Cohen's $f \approx 0.10$ (assuming 80% power and $\alpha = 0.05$). In the combined data, we were able to reliably detect even smaller effect sizes effect (Cohen's $f \approx 0.057$). We used interlocking quotas for age, gender, region, and income in Brazil and the United Kingdom; and age, gender, and region in India (on the Open Science Framework we provide the full breakdown). Quota targets were based on the online population and not the national population to avoid overrepresenting groups that are not connected to the internet—something that is particularly important in countries like India with relatively low internet penetration. The survey was designed by the authors and built using Qualtrics. We focus on these three countries because India is characterized by high levels of belief in conspiracy theories, the United Kingdom by particularly low levels, and Brazil falls in between (Kirk, 2022). Moreover, the authors possess in-depth expertise in each of these countries, including knowledge of the language and cultural context.

Design

Participants were presented with nine real Facebook posts. Six of the posts contained information that had been fact-checked and found to be false, while three contained true information. The false posts were all contemporary real-world examples of misinformation sourced from International Fact-Checking Network (IFCN)-accredited fact-checking organisations that partnered with Facebook to provide ratings that directly inform if and how Facebook labels content. The true posts were sourced from the Facebook pages of health authorities in each country. The posts were country-specific, and the posts displayed to participants in Brazil were in Portuguese; in the United Kingdom and India posts were in English. This makes comparisons between countries confounded, but it increases the real-world relevance of our findings, as the posts were actually circulating in the countries we investigated. Given the applied focus of our research, we chose to prioritize country-specific conclusions over cross-country comparisons.

Participants were randomly assigned to one of three conditions (see Figure 2). Participants in the Control (no correction) condition saw the nine posts (three true, six false) without any corrections.

Participants in the Correction condition saw the nine posts (three true, six false) with a user correction under four of the six false posts. Two false posts were left uncorrected to make the experiment more realistic (i.e., only some of the false posts people encounter on Facebook are likely to have been corrected by another user). We discuss these posts in Finding 5. Finally, the Correction with a link to a fact-check condition is identical to the Correction condition, except that the user correction was paired with a link to a fact check from a news organization (see panel C of Figure 2). At the end of the survey, all participants were debriefed, and all false posts were corrected (with links to fact checks).

In the United Kingdom, the fact checks were from the BBC or Reuters. In Brazil, they were from Aos Fatos, O Estado de S. Paulo, or Folha de S. Paulo. In India, they were from the BBC, The Quint, AFP, or Factly.

A) Control **B)** Correction **C)** Correction with a link

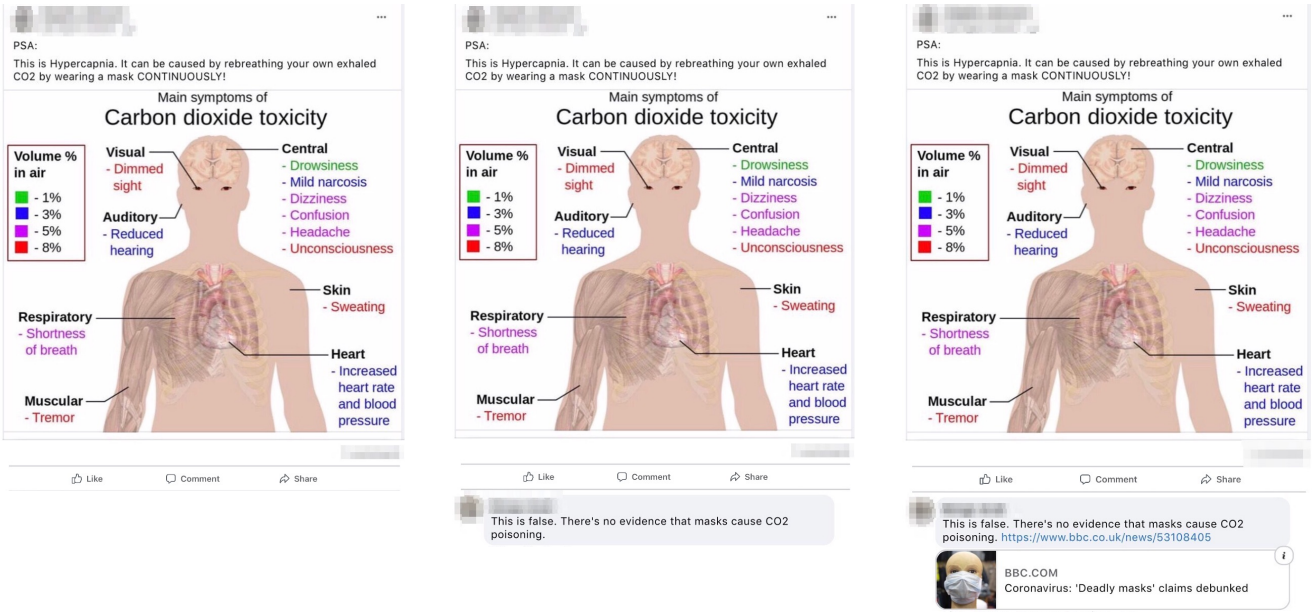

Figure 3. Examples of a post shown to participants in the United Kingdom in the Control condition, in which there is no correction (A), in the Correction condition, in which a user corrects the post (B), and Correction with a link to a fact check condition, in which a user corrects the post with a link to a fact check (C).

Measures

We first measured participants' demographics, trust, attitudes towards the news, news use, and belief in conspiracy theories (the full survey is available on the Open Science Framework). Before and after the treatment, participants passed an attention check (see Appendix B). Then, we measured the perceived accuracy of all posts using the following question, which contained a placeholder for a description of the claim relevant to the post being viewed: To the best of your knowledge, how accurate is the claim that <insert claim>? (0 = *not at all accurate*, 1 = *not very accurate*, 2 = *somewhat accurate*, 3 = *very accurate*). We measured participants' willingness to share the posts using the following question: How likely would you be to share this post on social media (e.g., on Facebook, Twitter, WhatsApp, etc.)? (0 = *not at all likely*, 1 = *not very likely*, 2 = *somewhat likely*, 3 = *very likely*).

We analyzed the data at the response level and used linear mixed-effects models. We included fixed effects for condition and post (and country in the combined data), and random intercepts for respondent ID to account for clustering (i.e., multiple answers per participant). We report the estimates (*b*). In

Appendix F, we show that the effect sizes remain unchanged when using OLS linear regression with clustered standard errors on participants and posts, but that all *p*-values are smaller, such that in Brazil, the effects of corrections are now significant in three instances.

Bibliography

Badrinathan, S., & Chauchard, S. (2024). "I don't think that's true, bro!" Social corrections of misinformation in India. *The International Journal of Press/Politics*, *29*(2), 394–416. https://doi.org/10.1177/19401612231158770

Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, *6*(10), 1372–1380. https://doi.org/10.1038/s41562-022-01388-6

Blair, R. A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., & Stainfield, C. J. (2024). Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology*, *55*, 101732. https://doi.org/10.1016/j.copsyc.2023.101732

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication*, *33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

Bode, L., Vraga, E. K., & Tang, R. (2024). User correction. *Current Opinion in Psychology*, *56*, 101786. https://www.sciencedirect.com/science/article/pii/S2352250X23002312

Brashier, N. M. (2024). Fighting misinformation among the most vulnerable users. *Current Opinion in Psychology*, *57*, 101813. https://doi.org/10.1016/j.copsyc.2024.101813

Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in Psychology*, *4*, 279. https://doi.org/10.3389/fpsyg.2013.00279

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature*, *630*(8015), 45–53. https://doi.org/10.1038/s41586-024-07417-w

Coppock, A. (2023). *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226821832.001.0001

Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus*, *3*(7), pgae217. https://academic.oup.com/pnasnexus/advance-article-abstract/doi/10.1093/pnasnexus/pgae217/7686087

Graham, M. H., & and Porter, E. V. (2025). Increasing demand for fact-checking. *Political Communication*, *42*(2), 325–348. https://doi.org/10.1080/10584609.2024.2395859

Guess, A., McGregor, S., Pennycook, G., & Rand, D. (2024). Unbundling digital media literacy tips: Results from two experiments. *OSF.* https://osf.io/u34fp/download

Guess, A., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

Kaplan, J. (2025, January 7). *More speech and fewer mistakes.* Meta Newsroom. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Kirk, I. (2022). *What conspiracy theories did people around the world believe in 2021?* YouGov. https://yougov.co.uk/topics/international/articles-reports/2022/02/08/what-conspiracy-theories-did-people-around-world-b

Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2024). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, *19*(2), 477–488. https://doi.org/10.1177/17456916231190388

Martel, C., & Rand, D. G. (2024). Fact-checker warning labels are effective even for those who distrust fact checkers. *Nature Human Behaviour*, *8*(10), 1957–1967. https://www.nature.com/articles/s41562-024-01973-x

McClain, C. (2021, May 4). *70% of U.S. social media users never or rarely post or share about political, social issues*. Pew Research Center. https://www.pewresearch.org/fact-tank/2021/05/04/70-of-u-s-social-media-users-never-or-rarely-post-or-share-about-political-social-issues/

Mercier, H. (2020). *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

Okong'o, J. (2025, April 9). *Meta is failing to stop dangerous disinformation in the world's most spoken languages.* Poynter. https://www.poynter.org/fact-checking/2025/meta-disinformation-non-english-languages/

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences, 25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 123–205). Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), e2104235118. https://doi.org/10.1073/pnas.2104235118

Porter, E., & Wood, T. J. (2024). Factual corrections: Concerns and current evidence. *Current Opinion in Psychology*, *55*, 101715. https://www.sciencedirect.com/science/article/pii/S2352250X23001604

Renault, T., Amariles, D. R., & Troussel, A. (2024). *Collaboratively adding context to social media posts reduces the sharing of false news*. arXiv. https://doi.org/10.48550/arXiv.2404.02803

Vraga, E. K., & Bode, L. (2021). Addressing COVID-19 misinformation on social media preemptively and responsively. *Emerging Infectious Diseases*, *27*(2), 396–403. https://doi.org/10.3201/eid2702.203139

Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2020). Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health Communication*, *36*(13), 1776–1784. https://doi.org/10.1080/10410236.2020.1794553

Yang, W., Wang, S., Peng, Z., Shi, C., Ma, X., & Yang, D. (2022). Know it to defeat it: Exploring health rumor characteristics and debunking efforts on Chinese social media during COVID-19 crisis. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*(1), 1157–1168. https://doi.org/10.1609/icwsm.v16i1.19366

Funding
This research was completed as part of the Oxford Martin Program on Misinformation, Science, and Media, funded by the Oxford Martin School and with further support from the BBC World Service as part of the Trusted News Initiative.

Competing interests
The authors declare no competing interests.

Ethics
The research protocol was approved by the University of Oxford Central University Research Ethics Committee. Participants provided their informed consent.

Data availability
All materials needed to replicate this study are available via the Harvard Dataverse: https://doi.org/10.7910/DVN/MGUSFA and https://osf.io/p6xjs/. The pre-registration is available at https://osf.io/mh5te.

Appendix A: Descriptive figure

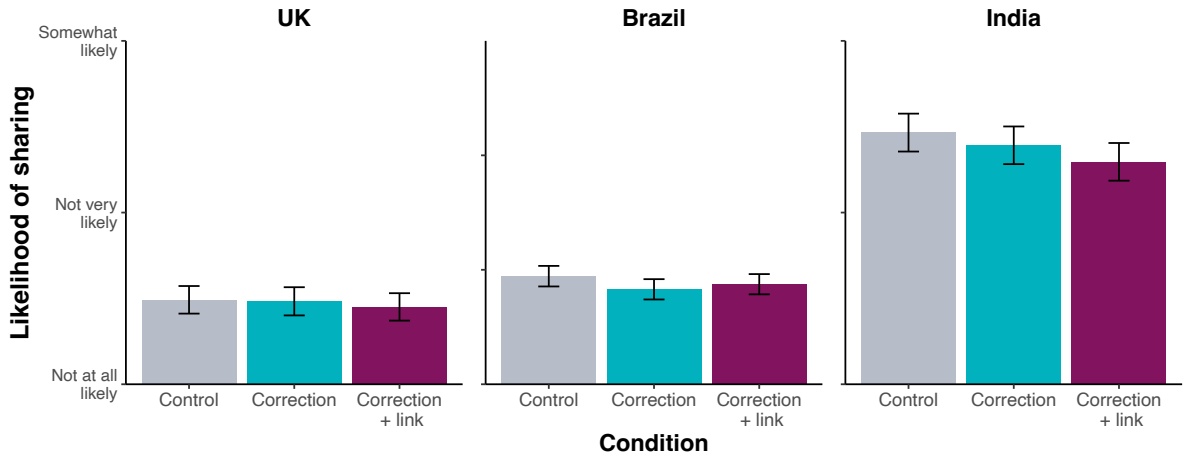

Figure A1. Bar plots representing participants' willingness to share false claims about COVID-19 in the Control condition (grey), Correction condition (blue), and Correction with a link to a fact check condition (purple). *The error bars represent the 95% confidence intervals. The full likelihood of sharing scale also includes a fourth point,* "very likely."

Appendix B: Excluding participants who failed the attention check

There were two attention checks: one pre-treatment: "The colour test is simple, when asked for your favourite colour you must choose the word puce below. Based on the text you read above, what colour have you been asked to choose?" where participants had to choose between 5 colors, including "puce" and one post-treatment: "The colour test is simple, when asked for your favourite colour you must choose the word brown below. Based on the text you read above, what colour have you been asked to choose?" where participants had to choose between 5 colors, including "brown" (and not "puce"). In India, 292 participants failed at least one attention check, compared to 75 in Brazil and 90 in the United Kingdom.

In the pre-registration, we said that "We will re-estimate and compare all of our analyses by dropping individuals who fail our attention checks." However, because it is not recommended to condition treatment effects on post-treatment variables, below we exclude only participants who failed the pre-treatment attention check.

We see that the effect sizes are similar when excluding participants who failed the attention check; however, the effects of corrections with a link in India are no longer significant, while the effect of correction without a link in India is significant. Beyond small differences in *p*-values (differences that are themselves not significant), these findings point in the direction that short user corrections may be slightly effective, but the effects are so small that very large sample sizes are needed to reliably detect them and that the added value of the link to fact checks is likely even smaller.

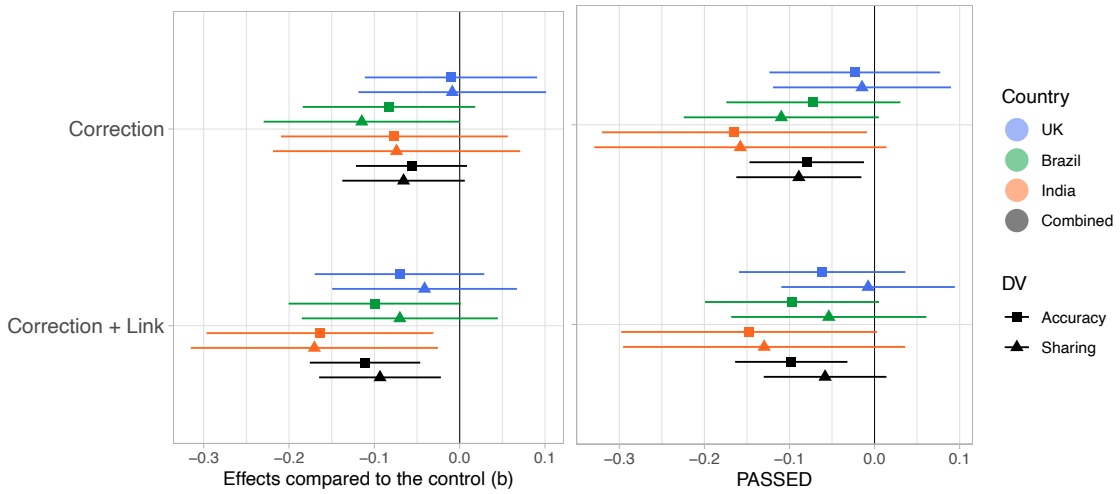

***Figure B1. Effects of the corrections compared to the control on the 4-point scale (b)**. In the left panel, we do not exclude participants who failed the attention check, while on the right panel, we exclude participants who failed the attention check.*

Appendix C: Determinants of belief in false COVID-19 and willingness to share it

Across countries, participants with higher levels of conspiracy ideation (b_{belief} = 0.17, $b_{sharing}$ = 0.17, p < .001), higher trust in social media (b_{belief} = 0.15, $b_{sharing}$ = 0.22, p < .001), higher trust in ordinary people (b_{belief} = 0.13, $b_{sharing}$ = 0.13, p < .001) and lower trust in scientists (b_{belief} = -0.14, $b_{sharing}$ = -0.12, p < .001) were more likely to believe and share COVID-19 misinformation. Younger participants were also more likely to believe and share COVID-19 misinformation (b_{belief} = -0.03, $b_{sharing}$ = -0.04, p < .001), while education and gender had no statistically significant effects. These effects are consistent across countries, except that trust in news was positively associated with belief in misinformation in India (b_{belief} = 0.15, $b_{sharing}$ = 0.18, p < .001), whereas the opposite was true in Brazil (b_{belief} = -0.11, $b_{sharing}$ = - 0.08, p < .001).

Appendix D: Moderators

In this section, we report the moderating role of conspiracy ideation, trust in social media, and trust in the news on the effectiveness of corrections. In separate models, and for both accuracy judgments and sharing intentions, we interacted conspiracy ideation, trust in social media, and trust in the news, with treatment condition. Overall, we found that conspiracy ideation, trust in social media, and trust in news organizations did not significantly moderate the effectiveness of corrections on accuracy judgments and sharing intentions.

Conspiracist ideation was measured using the four-item scale developed by Brotherton et al. (2013) and used by Bode and Vraga (2018) in their study of user corrections to health misinformation on social media. Participants were asked to indicate their belief in four statements using a five-point scale ranging from -2 (*definitely not true*), through 0 (*not sure/can't decide*), to 2 (*definitely true*).

Trust in social media for coronavirus information and trust in news organisations were measured using a single item adapted from those discussed by Strömbäck et al. (2020). Specifically, participants were asked, "How much do you trust each of the following for news and information about coronavirus (COVID-19)?" where "social media" and "news organisations" could be scored from 0 (*not at all*) to 4 (*a great deal*).

Conspiracy ideation

Conspiracy ideation did not significantly moderate the effectiveness of corrections on accuracy judgments (no link: $p = .95$, link: $p = .14$). In none of the three countries did conspiracy ideation significantly moderate the effectiveness of corrections on accuracy judgments.

Conspiracy ideation did not significantly moderate the effectiveness of corrections on sharing intentions (no link: $p = .65$, link: $p = .24$). In none of the three countries did conspiracy ideation significantly moderate the effectiveness of corrections on sharing intentions.

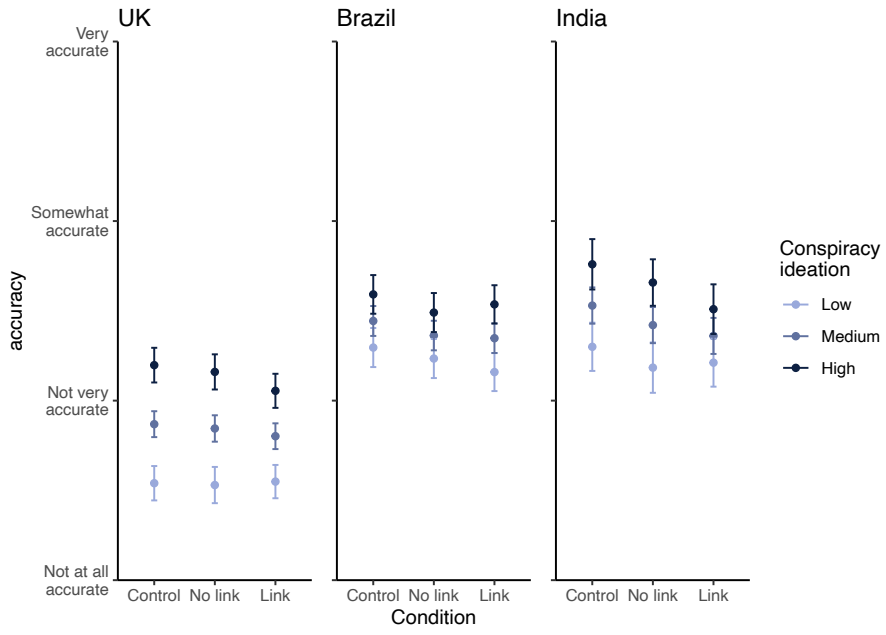

Figure D1. Moderating effect of conspiracy ideation on accuracy ratings. *The error bars represent the 95% confidence intervals.*

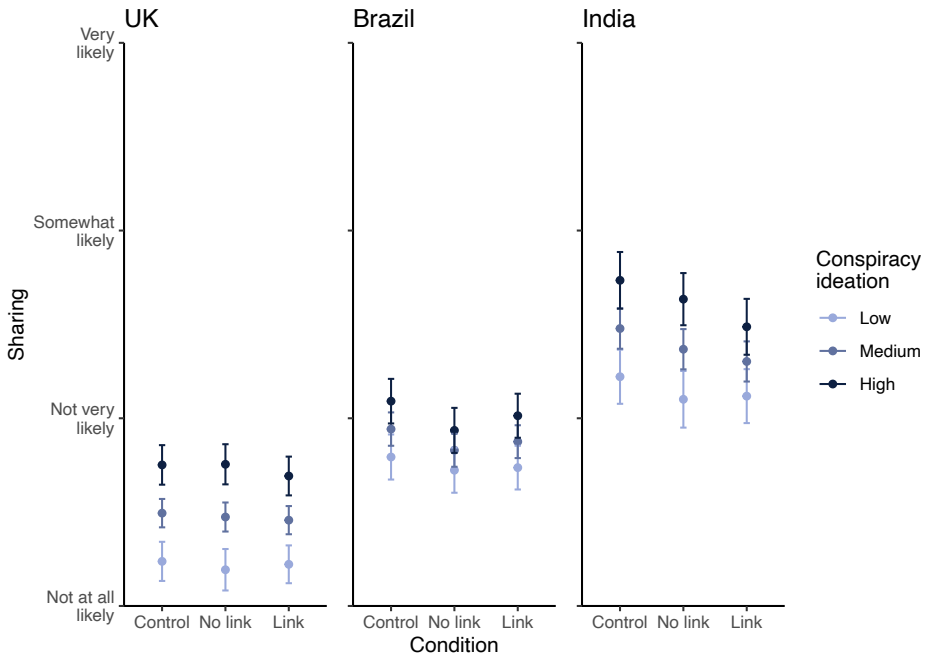

Figure D2. Moderating effect of conspiracy ideation on sharing intentions. *The error bars represent the 95% confidence intervals.*

Trust in social media

Trust in social media did not significantly moderate the effectiveness of corrections on accuracy judgments (no link: p = .78, link: p = .92). In none of the three countries did trust in social media significantly moderate the effectiveness of corrections on accuracy judgments. Trust in social media did not significantly moderate the effectiveness of corrections on sharing intentions (no link: p = .93, link: p = .40). In none of the three countries did trust in social media significantly moderate the effectiveness of corrections on sharing intentions.

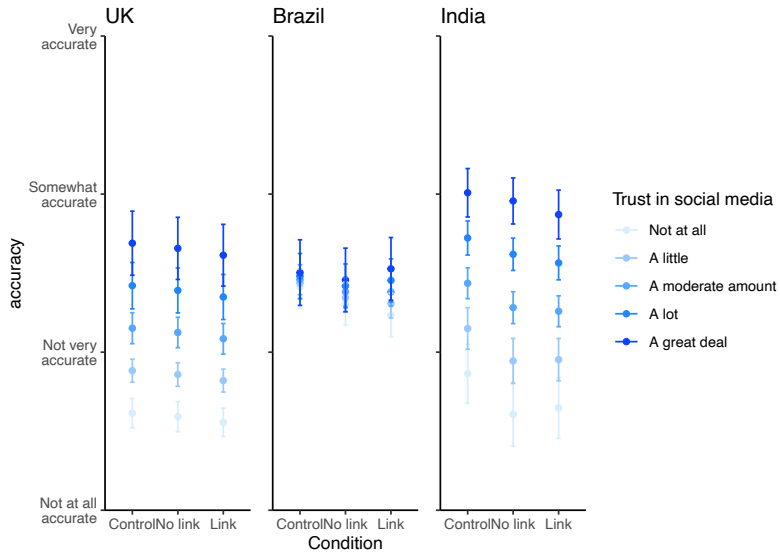

Figure D3. Moderating effect of trust in the news on accuracy ratings. *The error bars represent the 95% confidence intervals.*

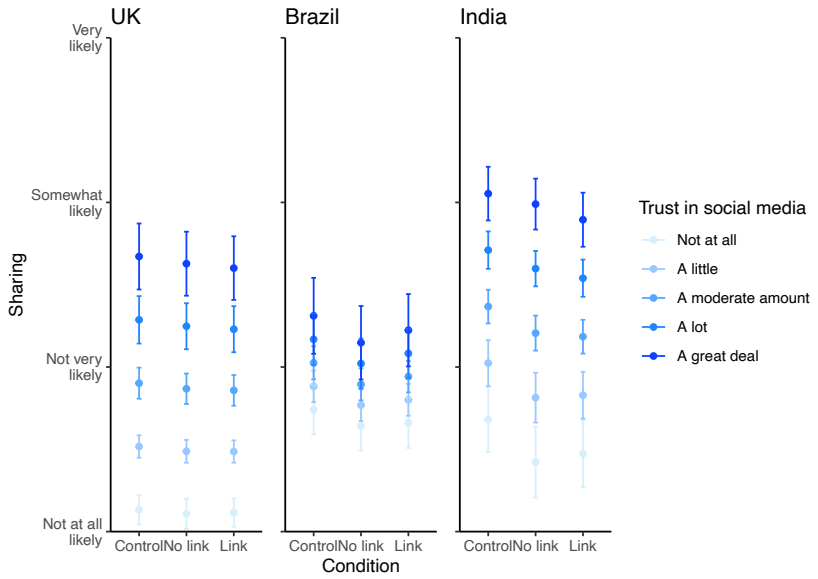

Figure D4. Moderating effect of trust in the news on sharing intentions. *The error bars represent the 95% confidence intervals.*

Trust in the news

Trust in news did not significantly moderate the effectiveness of corrections on accuracy judgments (no link: p = .22, link: p = .39). In none of the three countries did trust in news media significantly moderate the effectiveness of corrections on accuracy judgments. Trust in news did not significantly moderate the effectiveness of corrections on sharing intentions (no link: p = .31, link: p = .54). In none of the three countries did trust in news media significantly moderate the effectiveness of corrections on sharing intentions.

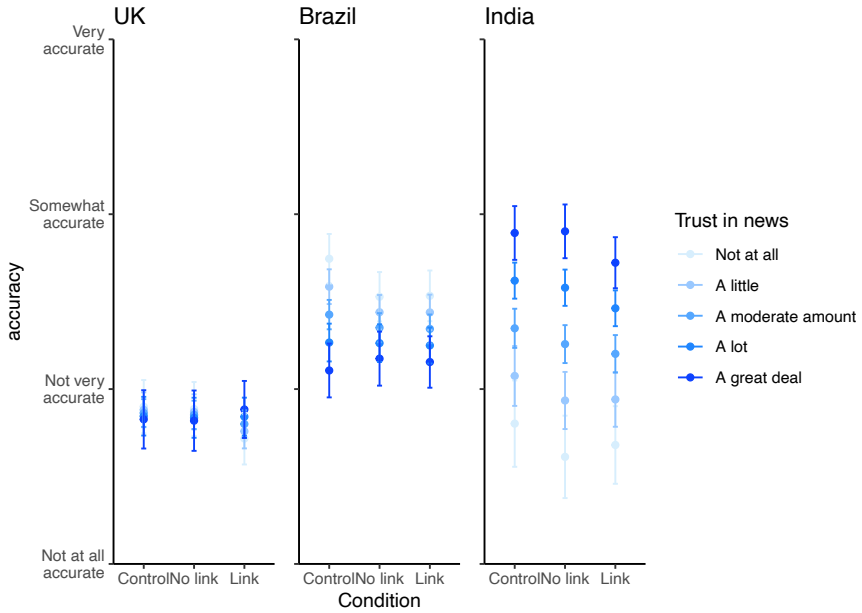

Figure D5. Moderating effect of trust in the news on accuracy ratings. *The error bars represent the 95% confidence intervals.*

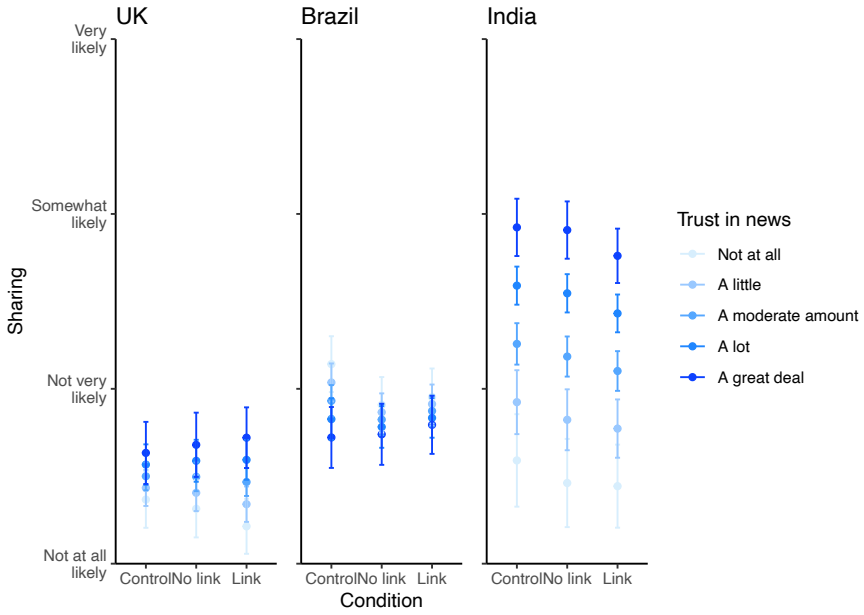

Figure D6. Moderating effect of trust in the news on sharing intentions. *The error bars represent the 95% confidence intervals.*

Appendix E: Descriptives on trust and political orientation

Country: United Kingdom

Table E1. *Mean and median levels of trust. "Generally speaking, to what extent do you trust information from the following media types in the United Kingdom?"* 0 = *not at all*, 1 = *a little*, 2 = *a moderate amount*, 3 = *a lot*, 4 = *a great deal*.

Source	*M*	*Mdn*
Newspapers	1.679	2
Television	2.204	2
Radio	2.022	2
Online news websites	1.803	2
Social media	1.037	1

Table E2. *Mean and median levels of trust in sources of news and information about COVID-19. "*How much do you trust each of the following for news and information about coronavirus (COVID-19)?" 0 = *not at all*, 1 = *a little*, 2 = *a moderate amount*, 3 = *a lot*, 4 = *a great deal*.

Source	*M*	*Mdn*
Scientists, doctors and other health experts	2.832	3
Ordinary people	1.366	1
News organisations	1.867	2
The government	1.980	2
Social media	0.941	1

Table E3. *Political orientation. Some people talk about "left," "right," and "centre" to describe parties and politicians. (Generally, socialist parties would be considered "left wing," whilst conservative parties would be considered "right wing"). With this in mind, where would you place yourself on the following scale?*

Ideology	Count
Very left-wing	30
Fairly left-wing	123
Slightly left of centre	121
Centre	283
Slightly right of centre	146
Fairly right-wing	88
Very right-wing	17
Don't know	192

Country: Brazil

Table E4. *Mean and median levels of trust. "Generally speaking, to what extent do you trust information from the following media types in Brazil?"* 0 = *not at all*, 1 = *a little*, 2 = *a moderate amount*, 3 = *a lot*, 4 = *a great deal*.

Source	*M*	*Mdn*
Newspapers	2.157	2
Television	1.989	2
Radio	2.102	2
Online news websites	1.872	2
Social media	1.547	2

Table E5. *Mean and median levels of trust in sources of news and information about COVID-19. "*How much do you trust each of the following for news and information about coronavirus (COVID-19)?"

0 = *not at all*, 1 = *a little*, 2 = *a moderate amount*, 3 = *a lot*, 4 = *a great deal*.

Source	*M*	*Mdn*
Scientists, doctors and other health experts	2.985	3
Ordinary people	1.253	1
News organisations	1.909	2
The government	1.377	1
Social media	1.480	2

Table E6. *Political orientation. Some people talk about "left," "right," and "centre" to describe parties and politicians. (Generally, socialist parties would be considered "left wing," whilst conservative parties would be considered "right wing"). With this in mind, where would you place yourself on the following scale?*

Ideology	Count
Very left-wing	64
Fairly left-wing	132
Slightly left of centre	78
Centre	198
Slightly right of centre	91
Fairly right-wing	116
Very right-wing	121
Don't know	200

Country: India

Table E7. *Mean and median levels of trust. "Generally speaking, to what extent do you trust information from the following media types in India?"* 0 = *not at all*, 1 = *a little*, 2 = *a moderate amount*, 3 = *a lot*, 4 = *a great deal*.

Source	*M*	*Mdn*
Newspapers	2.911	3.0
Television	2.760	3.0
Radio	2.263	2.0
Online news websites	2.688	3.0
Social media	2.510	2.5

Table E8. *Mean and median levels of trust in sources of news and information about COVID-19. "*How much do you trust each of the following for news and information about coronavirus (COVID-19)?" 0 = *not at all*, 1 = *a little*, 2 = *a moderate amount*, 3 = *a lot*, 4 = *a great deal*.

Source	*M*	*Mdn*
Scientists, doctors and other health experts	3.191	3
Ordinary people	2.227	2
News organisations	2.596	3
The government	2.832	3
Social media	2.359	2

Table E9. *Political orientation. Some people talk about "left," "right," and "centre" to describe parties and politicians. (Generally, socialist parties would be considered "left wing," whilst conservative parties would be considered "right wing"). With this in mind, where would you place yourself on the following scale?*

Ideology	Count
Very left-wing	68
Fairly left-wing	96
Slightly left of centre	82
Centre	422
Slightly right of centre	68
Fairly right-wing	91
Very right-wing	78
Don't know	95

Appendix F: Estimating treatment effect with OLS

Below, we estimate the effect of correction using OLS linear regression with clustered standard errors on participants and posts. We see that the estimates remain unchanged, but that all *p*-values are smaller in the OLS models. The digits in red indicate that there is a change in the OLS model compared to the linear mixed effect model.

Table F1. *Effect of the corrections on accuracy estimated with linear mixed effect models ("LMER" columns) and linear regressions ("OLS" column).*

	Country	Condition	*LMER b (p)*	*OLS b (p)*
Accuracy	UK	Link	-0.07 (.16)	-0.07 (.13)
		No link	-0.01 (.84)	-0.01 (.80)
	Brazil	Link	-0.10 (.054)	**-0.10 (.044)**
		No link	-0.08 (.11)	**-0.08 (.033)**
	India	Link	**-0.16 (.016)**	**-0.16 (.005)**
		No link	-0.08 (.26)	-0.08 (.21)

Table F2. *Effect of the corrections on sharing estimated with linear mixed effect models ("LMER" columns) and linear regressions ("OLS" column).*

	Country	Condition	*LMER b (p)*	*OLS b (p)*
Sharing	UK	Link	-0.04 (.46)	-0.04 (.39)
		No link	-0.01 (.88)	-0.01 (.85)
	Brazil	Link	-0.07 (.23)	-0.07 (.28)
		No link	-0.11 (.051)	**-0.11 (.038)**
	India	Link	**-0.17 (.021)**	**-0.17 (.009)**
		No link	-0.07 (.32)	-0.07 (.26)