How to Think About Whether Misinformation Interventions Work

Brian Guay *[1,2], Adam J. Berinsky[1], Gordon Pennycook[3], and David Rand[2]

[1]Department of Political Science, Massachusetts Institute of Technology
[2]Sloan School of Management, Massachusetts Institute of Technology
[3]Department of Psychology, Cornell University

August 10, 2023

Published in
Nature Human Behavior.

Progress in the burgeoning field of misinformation research requires some degree of consensus about what constitutes an effective intervention to combat misinformation. We differentiate between research designs that are used to evaluate interventions and recommend one that measures how well people discern between true and false content.

*brianmguay@gmail.com.

Growing concern about misinformation has spurred an explosion of research on who believes and shares false and misleading content[1–4], and what can be done about it[2;5]. Yet surprisingly little attention has been paid to the most fundamental prerequisites to answering these questions: how one should evaluate the efficacy of an intervention or the relative susceptibility of different groups to misinformation. Studies that purport to answer these same questions use different designs and analysis approaches, which inhibits our understanding of how to address the problem of misinformation.

For example, one common research design entails survey respondents rating a series of false (for example, as rated by professional fact-checkers) content on the likelihood that they believe it to be true and/or would share it[6–9]. Other studies ask respondents to rate a mix of false and true (that is, accurate) content[2;3;5;10]. Even among studies that include both false and true content, there is further variation in which outcomes scholars use to measure susceptibility to misinformation: some focus primarily on how much people believe or share the false content[11;12] and others focus on discernment — how much people believe or share the true content relative to the false content[2;5;10]. Using different research designs and outcomes can lead to conflicting conclusions about who is most likely to share false claims and which interventions are effective in combating them. Thus, for the field to move forward most effectively, it is necessary to bring coherence to the design and analysis approaches used.

We aim to rectify this issue by providing a unified framework for thinking about how to measure and operationalize susceptibility to misinformation. We consider past research in the context of normative claims that are (implicitly or explicitly) made about how citizens should engage with information. We argue that the appropriate normative claim — that citizens should maximize the accuracy of their beliefs and of the content that they share — requires (1) a design in which respondents rate a mix of both true and false content, and (2) an analysis that includes examining discernment between the two (rather than only examining false items).

Measuring ratings of false content

Misinformation studies often focus exclusively on how people interact with false content. This approach implies a normative claim that is at odds with the reality of the information environment on social media — namely, that users should not believe or share false content, but that whether they believe or share true content is inconsequential.

This normative claim is problematic for two reasons. First, after years of American politicians decrying unfavourable news coverage as fake and with trust in the US news media in recent years at an all-time low, disbelieving true news is an increasingly salient problem. Just as believing false content touting the benefits of ivermectin for treating COVID-19 is clearly problematic, so too is not believing true content about the benefits of masks or mRNA vaccines. Indeed, not believing true content is often synonymous with holding a false belief — in the case of COVID-19, not believing information about the effectiveness of vaccines can imply a belief that they are ineffective. Not sharing true content on social media may also have consequences. What users see on social media is largely determined by what their friends share. Although users do not necessarily have a responsibility to share all true content upon encountering it, sharing true content can crowd out false content.

Second, true news is far more prevalent than false news. Indeed, explicitly false content is rare on social media relative to true content and often originates from a small number of individuals[1;2]. Thus, studies that examine how people interact with only false content not only set up a highly unrealistic information environment but also overlook how people interact with the vast majority of content that they encounter.

In addition to these normative issues, there is also an important inferential issue with studies that use only false content: this design conflates the propensity to believe and share false content with the propensity to believe and share all content. A person may appear less likely to believe false content simply because they are less likely to believe all content — perhaps because they are distrusting of news in general, including true content[11;13]. Or they may share a great deal of false content because they are generally inclined to share in general (for example, particularly active social media

users). This design cannot distinguish these individuals from those who are specifically susceptible to believing or spreading false content per se.

This issue is particularly salient for studies that evaluate the efficacy of misinformation interventions, as interventions that are determined to be effective using only ratings of false content — but that have similar effects on true content — can actually do more harm than good. Such interventions may cause a general skepticism that disproportionately affects responses to true content, as this content is more prevalent on social media and thus is more frequently encountered. In fact, the goal of some disinformation campaigns could be to spread widespread disbelief and distrust, rather than promote a particular set of false beliefs. This is the approach that Russia is alleged to have taken during the 2016 US presidential election.

Assessing discernment

An alternative research design that addresses these limitations exposes participants to a mix of true and false content, and incorporates ratings of both into a measure of discernment. Discernment represents the extent to which a person believes or shares false content relative to true content. By capturing how individuals interact with both true and false content, discernment is more closely aligned with typical normative concerns over misinformation — that people cannot distinguish between true and false content. Discernment also reflects that benefits are derived not only from abstaining from believing and sharing false content, but also from believing and sharing true content.

As such, results of studies that use only false ratings and those that measure discernment can diverge in meaningful ways. We illustrate how using the hypothetical example of a study that examines the efficacy of a misinformation intervention, although the same logic applies to studies that compare beliefs in or sharing of false claims among nonexperimental groups (Democrats and Republicans, young versus old and so on). Figure 1 plots the effect of hypothetical treatments, each with different effects on belief in true (y axis) and false (x axis) content. Figure 1a determines the efficacy of an intervention using only ratings of false content, in which a treatment is considered

effective when it decreases belief in false content — regardless of its effect on true content. Notably, interventions in quadrants 2 and 3 are all determined to be effective (that is, helpful) because they have a negative effect on believing (or sharing) false content, regardless of their effects on true content.

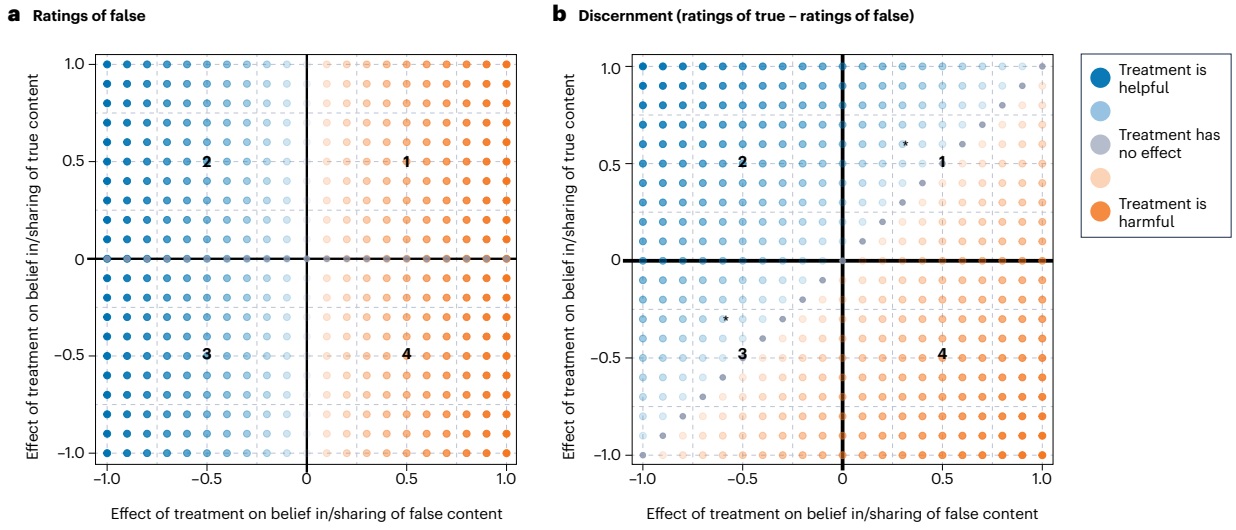

Fig. 1 | Using discernment vs. Ratings of Only False Content to Determine the Efficacy of Misinformation Interventions. a,b Efficacy of hypothetical misinformation interventions, as determined by ratings of only false content (**a**) and discernment between true and false content (**b**). In **a**, interventions are judged as effective if they have a negative effect on believing or sharing false content, regardless of their effect on ratings of true content. In **b**, however, interventions are judged as effective if they decrease belief in or sharing of false news more than they decrease belief in or sharing of true news. Whereas in **a** an intervention that decreases belief in all news (true and false) equally is judged as effective (that is, helpful), it is judged as having no effect in **b** because it does not improve a person's ability to distinguish between true and false content. The example of **b** also illustrates how different effects on belief and sharing of true and false content can result in identical effects on discernment: the two hypothetical interventions indicated by an asterisk in **b** have the same effect on discernment, despite the one in quadrant 1 in increasing belief in or sharing of trust and false content and the one in quadrant 3 decreasing belief in or sharing of true and false content.

Figure 1b shows the same data, but judges efficacy using discernment (which is jointly determined by the effect of the intervention on belief in true and false content). Interventions in quadrant 2 are still classified as effective as they both decrease belief in false content and increase belief in true content. However, now only half of the interventions in quadrant 3 are classified as effective — only those interventions that decrease belief in false content more than they decrease belief in true content. Likewise, half of the interventions in quadrant 1 are now classified as effective despite increasing belief in false content, because they increase belief in true content by a greater amount.

This example implicitly assumes that believing or sharing one piece of false content is as

normatively costly as believing or sharing one piece of true content is beneficial. Researchers should be explicit about this normative claim or else take a different normative stance and adjust their weighting accordingly (for example, believing true content may be upweighted relative to disbelieving false content given the far greater prevalence of true content). The key is to specify these claims explicitly in a preregistration before conducting the experiment to avoid adding additional experimenter degrees of freedom

Figure 1 also illustrates how different effects on belief and sharing of true and false content can result in identical effects on discernment. For instance, the two hypothetical interventions indicated by an asterisk in Fig. 1b have the same effect on discernment, despite the fact that the one in quadrant 1 increases belief in or sharing of true and false content and the one in quadrant 3 decreases belief in or sharing of true and false content.

In a supplementary analysis[14], we have re-analysed data from seven recent studies that asked respondents to rate true and false news content to illustrate the importance of belief and sharing discernment. Examples of interventions that decrease ratings of false headlines (that is, decrease belief or sharing) can have a positive effect on discernment either by increasing ratings of true headlines, having no effect on ratings of true headlines, having a negative effect (or smaller negative effect) on ratings of true headlines or having a positive effect on true headlines and no effect on false headlines. We also give examples of studies that significantly decrease ratings of false content despite having no effect on discernment given that it equally decreases ratings of true content.

Given that discernment is jointly determined by judgments of true and false content, it is critical to also examine its constituent parts to determine what drives the observed effect of discernment. Thus, a two-step approach is needed.

First, researchers should use discernment as the primary outcome of interest. Past work typically operationalizes discernment as the difference between average ratings of true versus false content (discernment = $\text{mean}_{true} - \text{mean}_{false}$). This is often done by modelling ratings of individual headlines with an interaction between dummy variables for veracity (true versus false) and group (for example, treatment versus control), typically using ordinary least squares with two-way standard

errors clustered on subject (that is, participant) and headline. There is a difference in discernment between groups when the interaction coefficient, which represents the difference-in-differences between ratings of true and false content in the treatment and control groups, is statistically significant. Importantly, the interaction used in this modelling approach provides the additive difference between true and false news across conditions, although other types of differences — such as multiplicative differences that capture relative differences between groups — can also be appropriate[15].

Second, given that discernment is jointly determined by judgments of true and false content, researchers should then separately examine effects on true and false content to determine what drives the effect (or lack of an effect) on discernment.

Optimizing research designs

The considerations and recommendations that we discuss here apply any time that researchers measure how much people believe or share misleading content. Most research on misinformation compares rates of believing or sharing misleading content across groups, whether those groups are randomly assigned — as in an experiment testing the efficacy of an intervention — or not. For instance, studies often compare rates of believing and sharing misleading content across political ideology[1], personality traits[11] and age[1].

Our primary objective is to guide researchers in choosing a research design that aligns with the intended goal of their study, rather than to prescribe a singular research design for all research on misinformation. Although we believe that for most studies on misinformation interventions the intended goal is to maximize the accuracy of the content people believe and share, this may not always be the case. For instance, an intervention may seek to reduce the overall amount of false content in the information environment regardless of the effect on true content. Likewise, an intervention may be intended to decrease belief in false news regardless of whether it decreases belief in true news as well. Thus, explicitly addressing and formalizing these goals enables researchers to preregister the research design and approach to analysing the results that most closely align with

their stated goals.

References

1. Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

2. Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545, 2020.

3. Benjamin A Lyons, Jacob M Montgomery, Andrew M Guess, Brendan Nyhan, and Jason Reifler. Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, 118(23):e2019527118, 2021.

4. Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, 115(3):999–1015, 2021.

5. Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780, 2020.

6. Fabian Zimmermann and Matthias Kohring. Mistrust, disinforming news, and vote choice: A panel survey on the origins and consequences of believing disinformation in the 2017 german parliamentary election. *Political Communication*, 37(2):215–237, 2020.

7. Andrea Pereira, Elizabeth Harris, and Jay J Van Bavel. Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes & Intergroup Relations*, 26(1):24–47, 2023.

8. Daniel Halpern, Sebastián Valenzuela, James Katz, and Juan Pablo Miranda. From belief in conspiracy theories to trust in others: Which factors influence exposure, believing and sharing fake news. In *International conference on human-computer interaction*, pages 217–232. Springer, 2019.

9. Simge Andı and Jesper Akesson. Nudging away false news: Evidence from a social norms experiment. *Digital Journalism*, 9(1):106–125, 2020.

10. Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592 (7855):590–595, 2021.

11. M Asher Lawson and Hemant Kakkar. Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news. *Journal of Experimental Psychology: General*, 151(5):1154, 2022.

12. Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4):1073–1095, 2020.

13. Rakoen Maertens, Friedrich Götz, Claudia R Schneider, Jon Roozenbeek, John R Kerr, Stefan Stieger, William Patrick McClanahan III, Karly Drabot, and Sander van der Linden. The misinformation susceptibility test (mist): A psychometrically validated measure of news veracity discernment. 2021.

14. B Guay, AJ Berinsky, G Pennycook, and D Rand. How to think about whether misinformation interventions work. psyarxiv.

15. Tyler J VanderWeele and Mirjam J Knol. A tutorial on interaction. *Epidemiologic methods*, 3 (1):33–72, 2014.

Competing Interests

The authors declare no competing interests.

Additional Information

Peer review information *Nature Human Behaviour* thanks Sacha Altay and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.