How To Think About Whether Misinformation Interventions Work

Brian Guay[1], Adam J. Berinsky[2], Gordon Pennycook[3], and David Rand[4]

[1,2]Department of Political Science, Massachusetts Institute of Technology
[3]Hill/Levene Schools of Business, University of Regina
[4]Sloan School of Management, Massachusetts Institute of Technology

March 13, 2023

Revise & Resubmit at *Nature Human Behavior*

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

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 prerequisite to answering these questions: how should one evaluate the efficacy of an intervention or the relative susceptibility of different groups to misinformation? Studies purporting to answer the same question use different designs and analysis approaches, inhibiting our understanding of how to address the problem of misinformation.

For example, one common research design entails survey respondents rating a series of false (e.g., 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 (i.e., 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 employed.

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 the 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 (i) a design in which respondents rate a mix of both true and false content, and (ii) an analysis that includes examining discernment between the two (rather than just examining false items). An intervention that decreases belief or sharing of false content while having no effect (or a positive effect) on true content is unambiguously effective - but the efficacy of interventions that simultaneously decrease (or increase)

both true and false content is unclear. In these situations, efficacy is dependent on the relative magnitude of the treatment effect on each type of content, as well as normative prescriptions of the cost of believing/sharing false content relative to the benefit of believing/sharing true content. Determining the efficacy of these interventions requires data on both true and false content, and requires researchers to compare their effect on true versus false content (i.e., discernment).

Measuring Ratings of False Content

Misinformation studies often focus exclusively on how people interact with false content. Despite the intuitive appeal of this approach, it is poorly suited to the task of studying how individuals interact with false claims. Most importantly, its use 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 unfavorable news coverage as fake and with trust in the U.S. 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 are 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, not believing information about the effectiveness of vaccines is synonymous with believing they are ineffective. Not sharing true content on social media is also consequential, as what users see on social media is largely determined by what their friends share. While users do not have a responsibility to share all true content upon encountering it, sharing true content crowds 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 examining 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 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[11] (see also Maertens et al. 2021). Without measuring belief in true and false content, these two scenarios are observationally equivalent. Inversely, some individuals may share a great deal of both true and false content, indicating that they are generally inclined to share (e.g. particularly active social media users) rather than being specifically susceptible to spreading false content per se.

This issue is particularly salient for studies that evaluate the efficacy of misinformation interventions, since interventions 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. Due to the higher prevalence of true content on social media, interventions that reduce believing and sharing of both true and false content will more frequently affect perceptions of true content. An individual will occasionally encounter and, as a result of the intervention, be skeptical of false content, however, they will much more frequently encounter and be skeptical of true content. In fact, the goal of some disinformation campaigns (e.g., Russia's in 2016) is to spread widespread disbelief, rather than promote a particular set of false beliefs.

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, though the same logic applies to studies that compare belief/sharing of false claims among non-experimental groups (e.g., Democrats/Republicans, young vs. old, etc.). Figure 1 plots the effect of hypothetical treatments, each with different effects on belief in true (y-axis) and false (x-axis) content. Panel A determines the efficacy of an intervention using only ratings of false content, where 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 (i.e., helpful) because they have a negative effect on believing (or sharing) false content, regardless of their effects on true content.

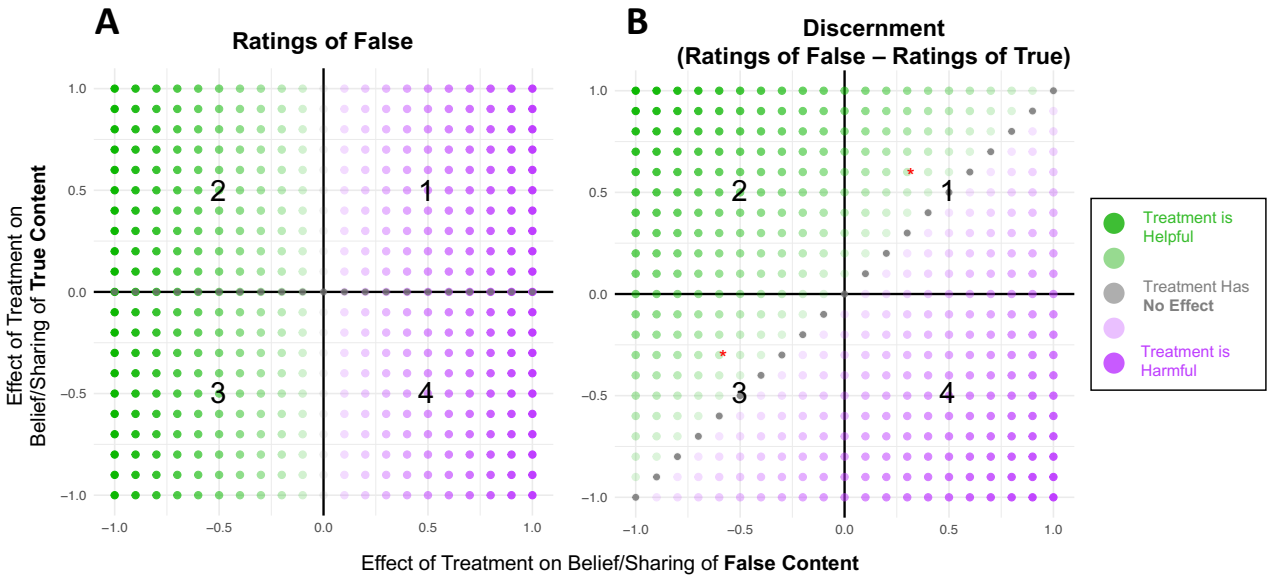

Fig. 1 | Using discernment vs. Ratings of Only False Content to Determine the Efficacy of Misinformation Interventions. Efficacy of hypothetical misinformation interventions, as determined by ratings of only false content (Panel A) and discernment between true and false content (Panel B). In Panel A, interventions are judged as effective if they have a negative effect on believing/sharing false content, regardless of their effect on ratings of true content. In Panel B, however, interventions are judged as effective if they decrease belief/sharing of false news more than they decrease belief/sharing of true news. While in Panel A an intervention that decreases belief in all news (true and false) equally is judged as effective (i.e., helpful), it is judged as having no effect on Panel B because it does not improve a person's ability to distinguish between true and false content.

Panel B shows the same simulated data, but judges efficacy using discernment, which is jointly determined by the intervention's effect 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 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.

Note that this 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 (in a pre-registration prior to conducting the experiment, to avoid adding additional experimenter degrees of freedom). Researchers implicitly make these normative claims without any discussion when adopting any research design—for instance, when using ratings of false content only the assumed benefit of believing/sharing true content is zero. On the other hand, the benefit of 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 when choosing the appropriate research design for a study.

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 Panel B have the same effect on discernment, despite the one in quadrant 1 increasing belief/sharing of true and false content and the one in quadrant 3 decreasing belief/sharing of true and false content.

In the Online Appendix, we re-analyze 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 (i.e., decrease belief/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 (smaller) negative effect on ratings of true headlines, having a

positive effect on true headlines and no effect on false headlines. We also give examples of studies that significantly decreases 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 is driving the observed effect of discernment. Thus, a two-step approach is needed.

First, 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 = meantrue - meanfalse). This is often done by modeling ratings of individual headlines with an interaction between dummy variables for veracity (true vs. false) and group (e.g., treatment vs. control), typically using OLS with two-way standard errors clustered on subject (i.e., 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 modeling approach provides the additive difference between true and false news across conditions, though other types of differences—such as multiplicative differences that capture relative differences between groups—can also be appropriate[13]. Second, separately examining effects on true and false content to determine what is driving the effect (or lack of an effect) on discernment.

Discussion

Despite the growing number of interventions aimed at reducing online misinformation, there is widespread incoherence in how the efficacy of these interventions is established. We recommend that researchers have participants rate a combination of true and false content, and use discernment between the two as the primary outcome used to determine the efficacy of interventions. This approach avoids conflating the propensity to believe/share false content with the propensity to believe/share all content, recognizes the problematic nature of not believing/sharing true content,

and aligns with the normative claim that people should maximize the accuracy of what they believe and share.

Importantly, while our primary focus is on selecting the appropriate research design for testing the efficacy of misinformation interventions, the same considerations and recommendations apply anytime 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. While 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 allows researchers to preregister the research design and approach to analyzing the results that most closely align with their stated goals.

References

1. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, & D. Lazer. *Science*, **363**(6425): 374–378 (2019).

2. A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, & N. Sircar. *Proceedings of the National Academy of Sciences*, **117**(27):15536–15545 (2020).

3. B. A. Lyons, J. M. Montgomery, A. M. Guess, B. Nyhan, & J. Reifler. *Proceedings of the National Academy of Sciences*, **118**(23):e2019527118 (2021).

4. M. Osmundsen, A. Bor, P. B. Vahlstrup, A. Bechmann, & M. B. Petersen. *American Political Science Review*, **115**(3):999–1015 (2021).

5. G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, & D. G. Rand. *Psychological Science*, **31**(7): 770–780 (2020).

6. F. Zimmermann & M. Kohring. *Political Communication*, **37**(2):215–237 (2020).

7. A. Pereira, E. Harris, & J. J. Van Bavel. *Group Processes & Intergroup Relations*, **26**(1):24–47 (2023).

8. D. Halpern, S. Valenzuela, J. Katz, & J. P. Miranda. In *Social Computing and Social Media. Design, Human Behavior and Analytics*, 217–232 (2019).

9. S. Andı & J. Akesson. *Digital Journalism*, **9**(1):106–125 (2020).

10. G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, & D. G. Rand. *Nature*, **592** (7855):590–595 (2021).

11. M. A. Lawson & H. Kakkar. *Journal of Experimental Psychology: General*, **151**(5):1154 (2022).

12. K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan, et al. *Political Behavior*, **42**:1073–1095 (2020).

13. T. J. VanderWeele & M. J. Knol. *Epidemiologic Methods*, **3**(1):33–72 (2014).

Competing Interests

The authors declare no competing interests.

Online Appendix For:
How To Think About Whether Misinformation Interventions Work

Brian Guay[1], Adam J. Berinsky[2], Gordon Pennycook[3], and David Rand[4]

[1,2]Department of Political Science, Massachusetts Institute of Technology
[3]Hill/Levene Schools of Business, University of Regina
[4]Sloan School of Management, Massachusetts Institute of Technology

March 13, 2023

Illustration with Data From Previous Studies

To illustrate the importance of belief and sharing discernment, we re-analyze data from seven recent studies that asked respondents to rate true and false news content. Across these studies, there is heterogeneity in both the construct of interest (belief vs. sharing) and the research question being evaluated (efficacy of misinformation interventions vs. subgroup differences in susceptibility to false claims).

Figure 1 illustrates the importance of measuring ratings of both true and false content. Interventions that decrease ratings of false headlines (i.e., decrease belief/sharing) can have a positive effect on discernment either by increasing ratings of true headlines[1](Panel A), having no effect on ratings of true headlines (B. Guay et al. Panel B), or also having a (smaller) negative effect on ratings of true headlines[2](Panel C). In all cases, the treatment has a negative effect on false headlines and a positive effect on discernment, despite having very different effects on true headlines.

Figure A1 here.

Fig. A1 | Different Operationalizations Lead to Different Conclusions. Reanalysis of misinformation studies that measure belief in and sharing of both true and false content. Treatment effects on false content, true content, and discernment are plotted separately for each study, where positive coefficients indicate that the treatment increased the outcome. Standard errors are clustered at the respondent and headline level, and vertical lines represent 90% and 95% confidence intervals.

The importance of true headlines is also illustrated in Panel D,[3] where the positive effect of the treatment on true headlines drives increased discernment, despite having no effect on false headlines. Finally, Panel E features a study (A. Arechar) in which an intervention has no effect on discernment because it significantly decreases ratings of true and false content. Without measuring ratings of true headlines, interventions that decrease belief/sharing of all content are observationally equivalent to those that specifically target false content. Studies like these are not uncommon. Indeed, Maertens et al. (2021) show that playing a fake news game decreases belief in true and false content equally, and therefore has no effect on discernment. Similarly, Lawson and Kakkar[4] argue that the liberal-conservative gap in sharing false content is moderated by conscientiousness, with low conscientious conservatives sharing more false content than low conscientiousness liberals. However, an identical pattern emerges for sharing of true content, and there is therefore no moderating effect of conscientiousness on discernment.[5]

References

1. V. Capraro and T. Celadin. *Personality and Social Psychology Bulletin*, 01461672221117691 (2022).

2. G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, & D. G. Rand. *Nature*, **592** (7855):590–595 (2021).

3. G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, & D. G. Rand. *Psychological Science*, **31**(7): 770–780 (2020).

4. M. A. Lawson & H. Kakkar. *Journal of Experimental Psychology: General*, **151**(5):1154 (2022).

5. H. Lin, G. Pennycook, and D. G. Rand. *Cognition*, **230**:105312 (2023).