How To Think About Whether Misinformation Interventions Work

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

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

September 28, 2022

Abstract

Recent years have seen a proliferation of experiments seeking to combat misinformation. Yet there has been little consistency across studies in how the effectiveness of interventions is evaluated, which undermines the field's ability to identify efficacious strategies. We provide a framework for differentiating between common research designs on the basis of the normative claims they make about how people should interact with information. We recommend an approach that aligns with the normative claim that citizens should maximize the accuracy of the content they believe and share, which requires (i) a design that includes both true and false content, and (ii) an analysis that includes examining discernment between the two. Using data from recent misinformation studies, we show that using the wrong research design can lead to misleading conclusions about who is most likely to spread misinformation and how to stop it.

This working paper has not yet been peer reviewed

Keywords: misinformation, fake news, social media, methodology

*Brian Guay; brianmguay@gmail.com; Massachusetts Institute of Technology

Introduction

Growing concern about misinformation has spurred an explosion of research on who believes and shares false and misleading content (e.g., Grinberg et al., 2019; Guess et al., 2019; Osmundsen et al., 2021), and what can be done about it (e.g., Pennycook et al., 2021; Guess et al., 2020; Nyhan et al., 2020; Pennycook and Rand, 2021). This body of work aims to answer a range of questions about misinformation: What causes people to believe and share false content? Are those who share falsehoods merely confused about what is true or do they intentionally spread falsehoods? How effective are various interventions aimed at combating misinformation?

Surprisingly little attention, however, has been paid to the most fundamental prerequisite to answering any of 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 (e.g., Zimmermann and Kohring, 2020; Pereira et al., 2021; Halpern et al., 2019; Tsang, 2021; Pretus et al., 2022; Andı and Akesson, 2020). Other studies ask respondents to rate a mix of false and true (i.e., accurate) content (e.g., Guess et al., 2020; Lyons et al., 2021; Pennycook et al., 2020, 2021). Even among studies that include both false and true content, there is further variation in which outcomes scholars use to measure susceptibility to misinformation, with some focused primarily on how much people believe or share the false content (e.g., Lawson and Kakkar, 2022; Clayton et al., 2020) and others focused on discernment—how much people believe or share the true content *relative to* the false content (e.g., Guay et al., 2022; Guess et al., 2020; Pennycook et al., 2020, 2021). 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.

This paper aims to rectify this issue by providing a unified framework for measuring and operationalizing susceptibility to misinformation. We consider the full range of research designs and specify the normative claims that each implicitly makes 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. 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 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).

Furthermore, we differentiate between two different types of discernment: additive discernment and multiplicative discernment. Additive discernment - the standard discernment measure used in the field - captures how much more true content a person believes or shares than false content (i.e. true minus false). Using additive discernment, for example, a person who shares true news 20% of the time and false news 10 % of the time would be characterized as being 10 percentage points more likely to share true news than false news. Here, we point out that there is another way of thinking about discernment: how many times as likely is a person to believe or share true content compared to false content (i.e. true divided by false). Using this multiplicative discernment measure, for example, that same person would instead be characterized as being twice as likely to share true news as false news. These two measures can lead to meaningfully distinct conclusions. For example, an intervention may reduce this person's probability of sharing both true and false news by 5 percentage points, and using additive discernment one would conclude that the intervention had no effect on discernment ($.2 - .1 = .15 - .05$). Using multiplicative discernment, however, the intervention would have been successful: the person would have gone from being two times more likely to share true news than false ($.2/.1$) to three times more likely ($.15/.05$). We argue that multiplicative discernment is often a better reflection of an intervention's goals than additive discernment and provide guidance for how researchers should choose between them.

Finally, we re-analyze data from recent misinformation studies to illustrate the importance of measuring discernment and the difference between the two types of discernment. We demonstrate empirically how using different analysis approaches can lead to divergent conclusions about who believes and shares false claims and the efficacy of misinformation interventions.

Issues With Measuring Ratings of Only 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 politicians decrying unfavorable news coverage as fake and with trust in the news media in recent years at an all-time low (Brenan, 2021), 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 (Grinberg et al., 2019; Guess et al., 2020; Nikolov et al., 2020). 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 (Maertens et al., 2021; Lawson and Kakkar, 2022). Without measuring belief in true and false content, these two scenarios are observationally equivalent. Inversely, some individuals may share a lot 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 (Maertens et al., 2021) 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 is to spread widespread disbelief, rather than promote a particular set of false beliefs (Yablokov, 2022).

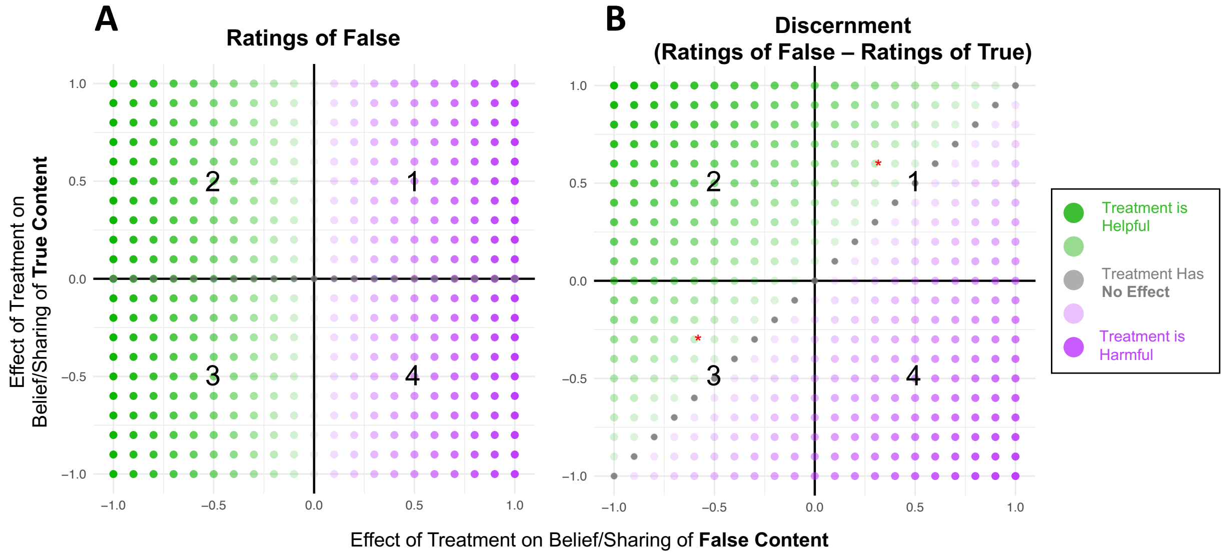
Accuracy and Sharing Discernment

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.

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

Figure 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.

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. 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. 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 a measure of discernment as the primary outcome of interest. Then decompose its constituent parts–separately examining effects on true and false content–to determine what is driving the effect (or lack of an effect) on discernment.[1]

Operationalizing Discernment

Past work typically operationalizes discernment as the difference between average ratings of true versus false content (discernment = $\text{mean}_{\text{true}}$ - $\text{mean}_{\text{false}}$). 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, and is therefore sometimes referred to as an additive interaction (VanderWeele and Knol, 2014).

[1]Note that in the absence of a significant interaction indicating discernment, finding a main effect on false content but not on true content (and vice versa) is *not* evidence of discernment. It the statistically significant difference between ratings of true and false content that indicates discernment, not merely the presence of an effect on one and not on the other.

$$\text{Additive Discernment} = (True_A - False_A) - (True_B - False_B) \quad (1)$$

where A and B refer to groups (e.g., treatment and control).

Panel A of Figure 2 illustrates an example of a hypothetical fake news intervention that increases additive discernment. In Panel A, additive discernment in the treatment group (0.9 - 0.3 = 0.6) is higher than in the control group (0.4 - 0.2 = 0.2), and the OLS coefficient representing the interaction between condition and headline veracity represents the difference between them (0.6 - 0.2 = 0.4), indicating that the treatment increases additive discernment.

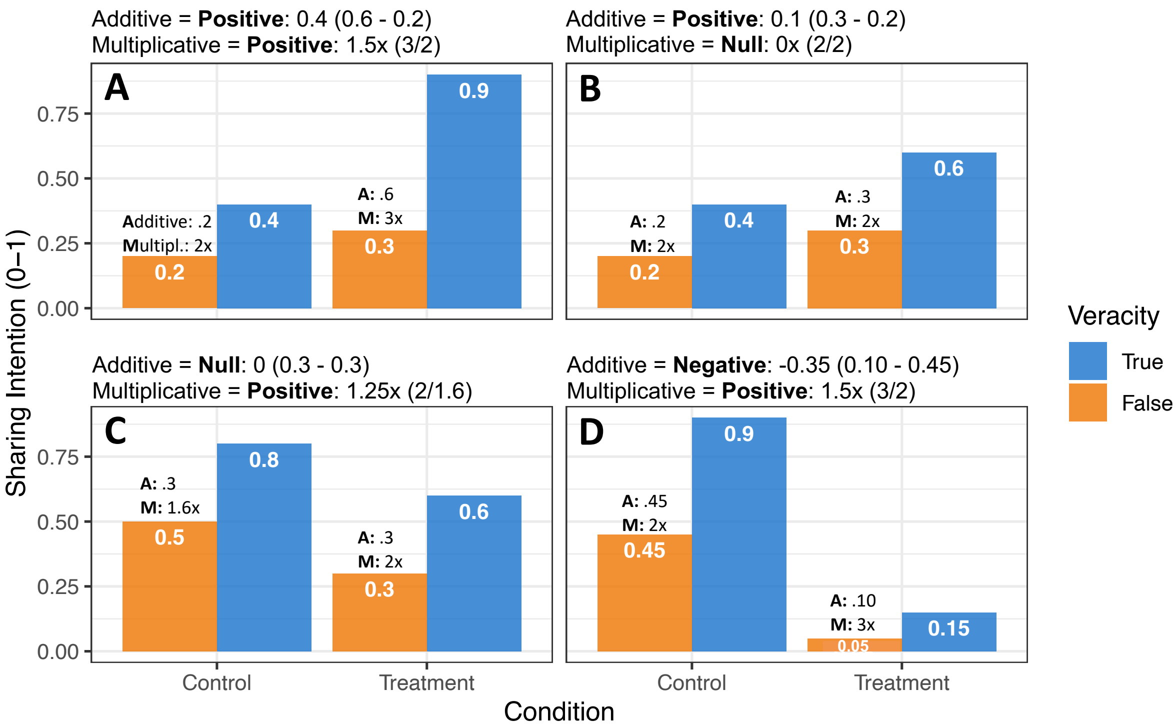
Additive differences are so commonplace in the social sciences that researchers rarely consider whether multiplicative differences are more appropriate. Multiplicative differences capture multiplicative, or relative, differences between groups (VanderWeele and Knol, 2014; Walter and Holford, 1978) and are calculated by computing a ratio of ratios—the ratio of true ratings to false ratings in one group (e.g., treatment group) to the same ratio in another group (e.g., control group).

$$\text{Multiplicative Discernment} = \frac{True_A/False_A}{True_B/False_B} \quad (2)$$

The intervention in Panel A that increases additive discernment also increases multiplicative discernment, given that the treatment group shares 3 times more true news than false news (0.9 / 0.3 = 3x) while the control group shares only 2 times more true news than false news (0.4 / 0.2 = 2x). The effect of treatment on multiplicative discernment is represented as the ratio of these two quantities (3 / 2 = 1.5).

Additive and multiplicative differences are not always in the same direction. For instance, Panel B illustrates a case in which the treatment has a positive effect on additive discernment but has zero effect on multiplicative discernment. Panel C illustrates the opposite case (zero effect on additive discernment and positive effect on multiplicative discernment), while Panel D illustrates a case in which the treatment effect is negative for additive discernment and

Figure 2: Additive vs. Multiplicative Discernment



Illustrative examples of the difference between additive and multiplicative discernment. In Panel A, the treatment increases both types of discernment. Additive discernment is calculated by subtracting the difference between ratings of true news and false news in the control condition (.4-.2 = .2) from the difference between ratings of true and false news in the treatment condition (.9-.3 = .6). This difference in differences (.6-.2 = .4) indicates that the treatment has a positive effect on additive discernment. Multiplicative discernment is calculated with ratios instead of differences: the ratio of true ratings to false ratings in the treatment group (.9/.3 = 3) divided by the ratio of true ratings to false ratings in the control groups (.4/.2=2). The resulting quotient (3/2 = 1.5) is the treatment effect on multiplicative discernment, indicating that multiplicative discernment is 1.5 times higher in the treatment group relative to the control group (multiplicative discernment that is greater than one indicates increased discernment). Additive and multiplicative discernment are calculated the same way in Panels B-D. In Panel C, the treatment has a positive effect on additive discernment, (.6-.3)-(.4-.2) = .1, but no effect on multiplicative discernment: (.6/.3)/(.4/.2) = 0. In Panel C, the treatment has no effect on additive discernment, (.8-.5)-(.6-.3) = 0, but has a positive effect on multiplicative discernment: the level of multiplicative discernment in the treatment groups is 1.25 times the level of multiplicative discernment in the control group ((.6/.3)/(.8/.5) = 1.25). Finally, in Panel D the treatment decreases additive discernment, (.15-.05)-(.9-.45)- = -0.35, but has a positive effect on multiplicative discernment ((.15/.05)/(.9/.45) = 1.5).

positive for multiplicative discernment.

These discrepancies between additive and multiplicative differences are the result of differences in overall sharing propensity. In Panel B, for instance, respondents in one condition are less likely to share any content (true or false) than the other, despite respondents in both conditions being twice as likely to share true content than false content. The difference between additive and multiplicative interaction is particularly important for evaluating the effect of misinformation interventions on discernment because it is not uncommon to see treatments that increase skepticism in general, decreasing people's tendency to believe and share *all* news (e.g., Maertens et al., 2021). This results in baseline differences in sharing and believing all news across conditions, but not necessarily improved ability to discern between true and false content. Similarly, baseline differences in general belief/sharing of content in general are common when assessing differences in discernment across subgroups in the population, such as political party (e.g., Republicans share more content in general; Guay et al., 2022) and age (e.g., older people share more content in general; Guess et al., 2019).

Additive differences are likely so ubiquitous in the social sciences because they can be calculated easily by including an interaction in a linear model estimated with Ordinary Least Squares (OLS). In the case of additive discernment, the coefficient for an interaction between headline veracity and experimental condition gives the difference-in-differences for ratings of true and false content across conditions (Equation 1). Fortunately, multiplicative differences can be calculated by specifying the same interaction, but in a model using a log link function rather than an identity link function. Since the log link puts the response variable on the multiplicative scale, the parameter estimates represent multiplicative effects.[2] Exponentiating these parameter estimates then provides an estimate of how much more discerning one group is compared to another, expressed in multiplicative terms (Equation 2). For instance, a value of 1 indicates that the treatment group is as discerning as the control group, a value of 1.5 indicates that the treatment group is 1.5 times as discerning as the control group, and a value of .5 indicates that the treatment group is half as discerning as the control group. Many models can be specified with a log-link function, including quasi-poison,

[2]Since $\exp(a + b) = \exp(a) * \exp(b)$.

negative binomial, gamma, and even (including the Gaussian model, i.e., normal model). R code to model additive and multiplicative discernment with and without clustered standard errors is available on Open Science Framework[3]

Is additive or multiplicative discernment more appropriate for judging the efficacy of misinformation interventions and assessing differences in discernment across subgroups? The answer depends on the researcher's normative claim about what the intervention should achieve, which has implications for how differences in baseline rates of belief or sharing are handled. By taking these baseline differences into account, multiplicative discernment reflects differences in the ability of individuals to discern between true and false news–the stated goal of many misinformation studies. As additive discernment does not take these baseline differences into account, it reflects differences in the total amount of true news shared related to false news. A researcher may, for instance, be concerned only with whether an intervention affects the total amount of false news that is shared online, regardless of whether baseline differences between the treatment and control group are driving this difference (e.g., as in Figure 2, Panel B).

To claim that an intervention is unambiguously effective, researchers should show evidence that it decreases both additive and multiplicative discernment. When researchers have a strong *a priori* reason to judge an intervention as effective if it increases only one type of discernment, however, that reason should be justified and the decision to model only one type of discernment should be pre-registered. Doing so favors analytical decisions that are grounded in theory rather than made post-hoc and improves our understanding of which interventions succeed in slowing the spread of misinformation.

[3]https://osf.io/ph5yv/?view$_o nly = 1294ddb906b24620aed8ec63cf85311e$

Re-analysis of Fake News Studies Using Belief and Sharing Discernment

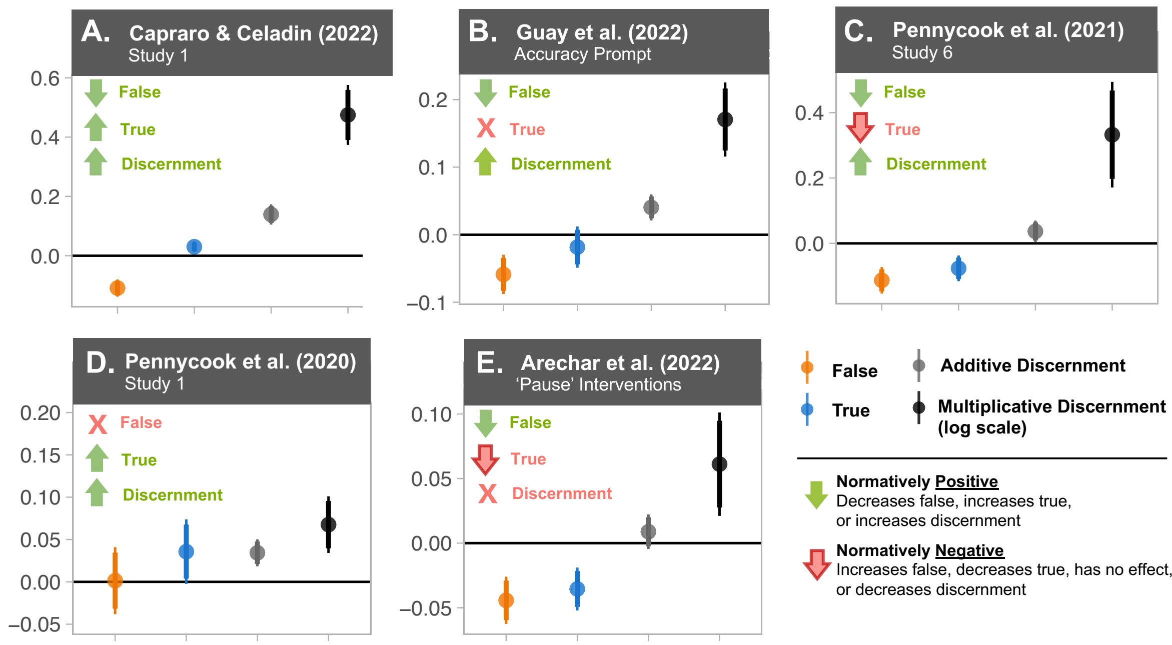
To illustrate the importance of belief and sharing discernment, we re-analyze data from seven recent studies that asked respondents to rate true and false news content. Across these studies, there is heterogeneity in both the construct of interest (belief vs. sharing) and the research question being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

Figure 3 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 (Capraro and Celadin, 2022, Panel A), having no effect on ratings of true headlines (Guay et al., 2022, Panel B), or also having a (smaller) negative effect on ratings of true headlines (Pennycook et al., 2021, 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.

The importance of true headlines is also illustrated in Panel D (Pennycook et al., 2020), 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 (Arechar et al. 2022) in which an intervention has no effect on additive discernment because it significantly decreases ratings of true and false content.[4] 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 (2022) argue that the liberal-conservative gap in sharing false content is moderated by conscientiousness, with low conscientious conservatives sharing more false content than

[4]As discussed below, the treatment has a positive effect on multiplicative discernment, despite having no effect on additive discernment.

Figure 3: Different Operationalizations Lead to Different Conclusions



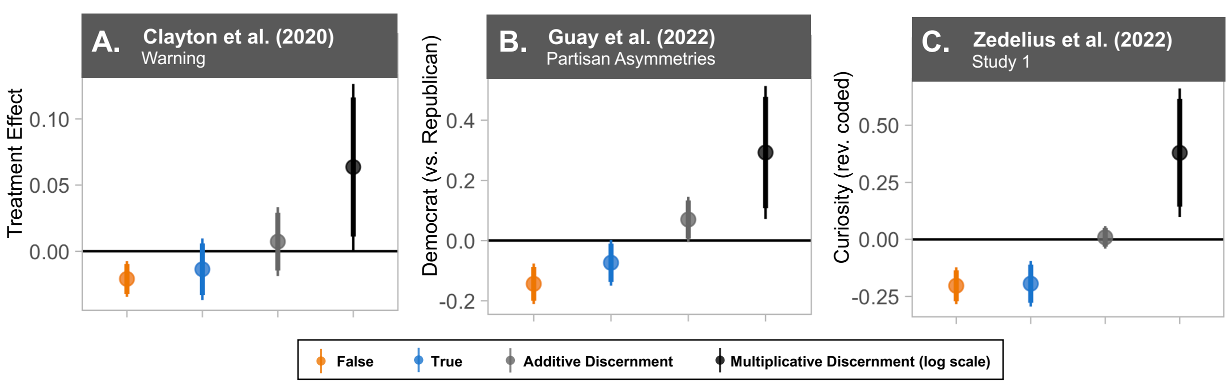
Reanalysis of misinformation studies that measure belief in and sharing of both true and false content. Treatment effects on false content, true content, additive discernment, and multiplicative discernment are plotted separately for each study, where positive coefficients indicate that the treatment increased the outcome. Coefficients representing multiplicative discernment are parameter estimates for the interaction between group (treatment vs. control) and veracity (true vs. false) from the Quasi-poisson models and are on the log scale. Exponentiating these coefficients provides an estimate of how much more likely one group is to share true (vs. false) news than another group. For instance, the parameter estimate for multiplicative discernment in Capraro and Celadin (2022) is 0.78 (log scale). Exponentiated, this quantity ($e^{0.78} = 2.18$) indicates that the treatment group has 2.18 times higher discernment than the control group. Standard errors are clustered at the respondent and headline level, and vertical lines represent 90% and 95% confidence intervals.

low conscientiousness liberals. However, an identical pattern emerges for sharing of true content, and there is therefore no moderating effect of conscientiousness on discernment (Lin and Pennycook, 2022).

Panel E also illustrates how additive and multiplicative discernment can differ: while the treatment has no effect on additive discernment, it has a positive effect on multiplicative discernment. Figure 4 illustrates more of these cases. Clayton et al.'s (2020) misinformation warning has no significant effect on additive discernment (p = .58), but has a significant effect on multiplicative discernment (p = .047). Similarly, Guay et al. (2022) find that

Democrats have higher levels of multiplicative discernment (p = .01) but not additive discernment (p = .07). Similar patterns are evidenced in data from Zedelius et al. (2022), which indicate that deprivation curiosity is associated with multiplicative (p = .01) but not additive discernment (p = .72). As discussed previously, discrepancies between additive and multiplicative discernment occur when there are differences in overall sharing rates across groups (e.g., treatment and control), and these differences are evident here. For instance, Republicans in Guay et al. (2022) share more of all types of news (true and false) on average than Democrats (Republicans = 0.41, Democrats: 0.30; difference = .11, p = .001), people with higher levels of deprivation curiosity in Zedelius et al. (2022) share news of any kind than those with lower levels (high curiosity = .50, low curiosity = 0.44; difference = .07, p = .01), and people in the Clayton et al. (2020) treatment condition believe more than people in the control condition (treatment = .46, control = .49; difference = .02, p = .001).

Figure 4: Additive and Multiplicative Discernment Can Lead to Different Conclusions



Examples of recent studies with different findings for additive and multiplicative discernment. Standard errors are clustered at the respondent and headline level, and vertical lines represent 90% and 95% confidence intervals. Once again, coefficients representing multiplicative discernment are on the log scale, and once exponentiated indicate the degree to which one group is more discerning than the other. For instance, in Clayton et al. (2020), people who received a misinformation warning had 1.06 times higher discernment than those who did not ($e^{0.06} = 1.06$); in Guay et al. (2022), Democrats have 1.34 higher discernment than Republicans ($e^{0.29} = 1.34$); and in Zedelius et al. (2022), people with low deprivation curiosity have 1.46 timer higher discernment than those with high deprivation curiosity ($e^{0.38} = 1.46$)

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 provide a detailed discussion of common research designs and a clear set of recommendations for how to measure who believes and shares misinformation, and when interventions work. Specifically, 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 false content with the propensity to believe all content, recognizes the problematic nature of not believing true content, and aligns with the normative claim that people should maximize the accuracy of what they believe and share. We also introduce a multiplicative operationalization of discernment, which can capture the ability to distinguish between true and false content better than additive discernment because it accounts for baseline differences in how likely people are to believe or share news of any kind.

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 (Grinberg et al., 2019; Guay et al., 2022), personality traits (Lawson and Kakkar, 2022), and age (Guess et al., 2019).

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. This paper provides a framework for explicitly discussing and formalizing these goals, allowing researchers to pre-register the research design and approach to analyzing the results that most closely aligns with their stated goals.

References

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

Brenan, M. (2021). Americans' trust in media dips to second lowest on record. *Gallup. https://news. gallup. com/poll/355526/americans-trust-media-dips-second-lowest-record. aspx*.

Capraro, V. and T. Celadin (2022). "i think this news is accurate": Endorsing accuracy decreases the sharing of fake news and increases the sharing of real news. *Personality and Social Psychology Bulletin*, 01461672221117691.

Clayton, K., S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan, et al. (2020). 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.

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer (2019). Fake news on twitter during the 2016 us presidential election. *Science 363*(6425), 374–378.

Guay, B., G. Pennycook, D. Rand, et al. (2022). Examining partisan asymmetries in fake news sharing and the efficacy of accuracy prompt interventions.

Guess, A., J. Nagler, and J. Tucker (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances 5*(1), eaau4586.

Guess, A. M., M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, and N. Sircar (2020). 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.

Halpern, D., S. Valenzuela, J. Katz, and J. P. Miranda (2019). 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*, pp. 217–232. Springer.

Lawson, M. A. and H. Kakkar (2022). 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.

Lin, Hause, H. R. D. G. and G. Pennycook (2022). Conscientiousness does not moderate the association between political orientation and susceptibility to fake news sharing while actively open-minded thinking does. *Under Review*.

Lyons, B. A., J. M. Montgomery, A. M. Guess, B. Nyhan, and J. Reifler (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences 118*(23), e2019527118.

Maertens, R., F. Götz, C. R. Schneider, J. Roozenbeek, J. R. Kerr, S. Stieger, W. P. McClanahan III, K. Drabot, and S. van der Linden (2021). The misinformation susceptibility test (mist): A psychometrically validated measure of news veracity discernment.

Nikolov, D., A. Flammini, and F. Menczer (2020). Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *arXiv preprint arXiv:2010.01462*.

Nyhan, B., E. Porter, J. Reifler, and T. J. Wood (2020). Taking fact-checks literally but not seriously? the effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior 42*(3), 939–960.

Osmundsen, M., A. Bor, P. B. Vahlstrup, A. Bechmann, and M. B. Petersen (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review 115*(3), 999–1015.

Pennycook, G., Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand (2021). Shifting attention to accuracy can reduce misinformation online. *Nature 592*(7855), 590–595.

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

Pennycook, G. and D. G. Rand (2021). The psychology of fake news. *Trends in cognitive sciences 25*(5), 388–402.

Pereira, A., E. Harris, and J. J. Van Bavel (2021). Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes & Intergroup Relations*, 13684302211030004.

Pretus, C., C. Servin-Barthet, E. Harris, W. Brady, O. Vilarroya, and J. Van Bavel (2022). The role of political devotion in sharing partisan misinformation.

Tsang, S. J. (2021). Motivated fake news perception: The impact of news sources and policy support on audiences' assessment of news fakeness. *Journalism & Mass Communication Quarterly 98*(4), 1059–1077.

VanderWeele, T. J. and M. J. Knol (2014). A tutorial on interaction. *Epidemiologic methods 3*(1), 33–72.

Walter, S. and T. Holford (1978). Additive, multiplicative, and other models for disease risks. *American Journal of Epidemiology 108*(5), 341–346.

Yablokov, I. (2022). Russian disinformation finds fertile ground in the west. *Nature Human Behaviour*, 1–2.

Zedelius, C. M., M. E. Gross, and J. W. Schooler (2022). Inquisitive but not discerning: Deprivation curiosity is associated with excessive openness to inaccurate information. *Journal of Research in Personality 98*, 104227.

Zimmermann, F. and M. Kohring (2020). 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.