advances.in/psychology

July 27, 2024
Doi: 10.56296/aip00019
ISSN: 2976-937X

PERSPECTIVE | SPECIAL ISSUE: FROM VULNERABILITY TO VIGILANCE

Beyond the headlines: On the efficacy and effectiveness of misinformation interventions

Jon Roozenbeek[1,2*], Miriam Remshard[2], & Yara Kyrychenko[2]

Received: April 30, 2024 | **Accepted:** July 24, 2024 | **Published:** July 27, 2024 | **Edited by:** Jonas R. Kunst

[1] Department of War Studies, King's College London, London, United Kingdom. [2] Department of Psychology, University of Cambridge, Cambridge, United Kingdom. *Please address correspondence to Jon Roozenbeek, jjr51@cam.ac.uk, Department of War Studies, King's College London, Strand, London, WC2R 2LS, United Kingdom.

Research into how to best counter misinformation has enjoyed a great deal of popularity, but a discussion about how efficacy (successful lab studies) translates to effectiveness (real-world impact) is lacking. Lab studies have shown that many types of misinformation interventions are efficacious at achieving their intended outcomes (e.g., improving "discernment", or the ability to distinguish true from false information). However, drawing on implementation science, we identify six challenges facing misinformation interventions research: (1) an overabundance of lab research and a lack of field studies; (2) the presence of testing effects, which impede intervention longevity and scalability; (3) modest effects for small fractions of relevant audiences; (4) a reliance on item evaluation tasks (e.g., rating a series of headlines as true or false) as the primary efficacy measure of interest; (5) low replicability in the Global South and a lack of audience-tailored interventions; and (6) an underappreciation of potential unintended consequences of intervention implementation. We argue that it is time to look beyond item task performance as the primary outcome measure and to elevate both real-world outcomes and alternative measures of effectiveness (e.g., intervention attractiveness or user uptake) as equally important ways of assessing "what works". We provide practical recommendations for addressing each challenge and improving intervention effectiveness.

Keywords: misinformation, implementation science, interventions, efficacy, effectiveness

1. INTRODUCTION

The spread of misinformation has become a keystone political and scientific challenge (Ecker et al., 2024), and has thus brought about a bustling field of scientific inquiry. Research into countering misinformation took flight around 2016 and has soared high ever since, with hundreds if not thousands of intervention studies yielding sometimes middling but often impressive results (Fazio et al., 2024; Guess et al., 2023; Lu et al., 2023). Such interventions either occur at the system level (for instance changing recommender algorithms; see Guess et al., 2023) or at the individual level (Roozenbeek et al., 2023). Individual-level interventions can, in turn, be divided into three categories (Kozyreva et al., 2024): *nudges* (aimed at behavioural change, for example, improving the quality of people's news sharing decisions), *boosts* (which foster competencies such as critical thinking or media/digital literacy, or seek to build new ones), and *refutation strategies* (tools to refute or correct misinformation, such as debunking and content labelling). Recently, Kozyreva et al. (2024) put together a "toolbox" of interventions from 81 scientific papers, along with information on how these interventions were tested and validated. As the field of misinformation research reaches maturity, a critical assessment of the research on interventions to counter misinformation is warranted.

Here, we outline key challenges and recommendations for improving the effectiveness of misinformation interventions. We draw on implementation science, or "the scientific study of methods to promote the systematic uptake of research findings and other evidence-based practices into routine practice" (Eccles & Mittman, 2006). A key distinction within this domain is between *efficacy* ("the performance of an intervention under ideal and controlled circumstances") and *effectiveness* ("performance under 'real-world' conditions"; Singal et al., 2014). Implementation scientists argue that establishing efficacy in lab studies does not guarantee that effectiveness is achieved, as an intervention's uptake and longevity depend on contextual factors and barriers (Bauer & Kirchner, 2020; Bauer et al., 2015). Both in the field of medicine (Hobby et al., 2022) and in the social and behavioural sciences (Brauer, 2024; Done et al., 2011; Hagger & Weed, 2019), the question of whether interventions "work" in the real world has been an important and often contentious topic of discussion.

Broadly speaking, misinformation interventions have been efficacious at achieving their desired goals, particularly when looking at results from (online) panel and survey studies (Lu et al., 2023; Pennycook & Rand, 2022). In a recent mega-study with nine different types of prominent misinformation interventions such as accuracy nudges, debunking, media literacy tips, and psychological inoculation, Fazio et al. (2024) found that "a wide variety of interventions can improve discernment between true versus false or misleading information", and conclude that "[countering] misinformation belief and sharing is a goal that is accomplishable through multiple strategies targeting different psychological mechanisms".

The prevailing methodological standard in this field of research has been to assess the efficacy of (individual-level) interventions through item evaluation tasks (Fazio et al., 2024; Lu et al., 2023). Such tasks typically consist of participants rating a series of social media posts or news headlines on a scale assessing either accuracy judgments or sharing intentions (e.g., "*To the best of your knowledge, is this headline accurate?*", with 1 being "entirely inaccurate" and 6 being "entirely accurate"). Although the field has moved towards a more standardised and validated measurement of misinformation veracity discernment (Maertens et al., 2024a), item evaluation task performance continues to take primacy in determining intervention efficacy (see Fazio et al., 2024), whereas other potentially equally relevant factors (such as entertainment value or user retention potential) are

considered of secondary importance or, as is often the case, not measured at all.

More broadly, how this type of efficacy testing translates to relevant real-world effectiveness (such as the change in the likelihood of someone believing a piece of misinformation they come across online or the frequency of sharing misinformation with others) relies on a set of assumptions that are often left implicit and are rarely tested in experimental studies. Furthermore, "what works" in some places often translates poorly across cultural, linguistic, and socio-economic contexts (Badrinathan & Chauchard, 2023): successful intervention studies conducted in the West more often than not fail to replicate in the Global South (Guess et al., 2020; Harjani et al., 2023), and researchers often continue to underestimate the importance of cultural knowledge and expertise when designing both interventions and the studies used to assess their effectiveness (Brashier, 2024).

Here, we argue that the lack of a discussion about efficacy versus effectiveness, or what works for whom when it comes to countering misinformation, has led to challenges translating lab research into real-world change. These challenges may be alleviated by designing studies with both real-world effectiveness and individual- and group-level contexts in mind. We note that the purpose of this article is not to cast doubt on the usefulness of (research into) misinformation interventions as a whole, nor do we argue that (individual-level) interventions are ineffective altogether (cf. Chater & Loewenstein, 2023). Rather, our goal is to make explicit what we believe are important assumptions and potential shortcomings that underlie this bustling research field, with a view to moving it forward. More generally, we hope that our discussion of the challenges of implementing misinformation interventions provides insights for implementation scientists working in other fields of inquiry.

2. CHALLENGES

2.1 Challenge #1: Prioritising Efficacy Over Effectiveness

Misinformation researchers assume that the item evaluation tasks used to test interventions in lab studies translate to real life in a meaningful way. For instance, if a lab-based intervention study yields improved veracity discernment (taken to mean one's ability to distinguish between true and false and/or misleading information; see Maertens et al., 2024a; Pennycook & Rand, 2019), then this ought to mean that a good number of people who interacted with the intervention as intended are more likely than before to correctly evaluate information they see on social media or elsewhere as true or false (or misleading, and so on). Similarly, improved "sharing discernment" (meaning the "quality of news sharing decisions", Pennycook et al., 2020; Roozenbeek et al., 2021) ought to yield higher-quality content sharing in the real world, with individuals impacted by the intervention sharing less unwanted content or more high-quality content, or both. However, there are several important knowledge gaps that limit our confidence that interventions reliably and predictably lead to such outcomes.

First, lab studies are conducted in optimised environments: participants are paid to pay attention, and if they fail an attention check their responses are excluded. We know from studies comparing results from lab experiments to those obtained in real-world environments (e.g., DellaVigna & Linos, 2022; Roozenbeek et al., 2022a) that effect sizes in field studies (in which conditions are far less optimal) are reduced substantially compared to what we find in the lab; a reduction by a factor of about six in the case of "nudge" interventions (DellaVigna & Linos, 2022)[1]. This indicates that the real-world effectiveness achieved by many misinformation interventions may be small if not

[1] Roozenbeek et al. (2022a) report comparable numbers in the context of an "inoculation" intervention implemented on YouTube.

negligible, especially for interventions that yield small effect sizes even in lab studies.

Second, field studies into how learning-based interventions affect people's *actual* (not self-reported) behaviour (e.g., what they engage with or share online post-intervention) are few and far between (Lees et al., 2023; see McPhedran et al., 2023, for a lab study with a relatively high degree of ecological validity), and in the case of behaviour-based interventions regularly yield either weak overall effects (Lin et al., 2024; Pennycook et al., 2021) or null results (e.g., Aslett et al., 2022). One study that looked at the impact of a learning-based intervention (a "Spot the Troll" quiz) on sharing behaviour on Twitter (Lees et al., 2023) showed that behavioural change, if it can be achieved at all, may be imprecise: rather than engaging less with only unwanted content, participants in the intervention condition showed lower rates of retweeting *any* content (not just the type of content targeted by the intervention) on Twitter/X for about a week compared to a control group, despite the intervention yielding improved veracity discernment in the lab.

Research assessing the impacts of misinformation interventions on *offline* behaviour has further highlighted the challenges of achieving real-world impact. For example, one study showed that, although correcting politicians' false statements led people to revise their beliefs about those claims, it did not alter their voting intentions (Swire et al., 2017)[2]. More promisingly, however, misinformation interventions in the context of climate change have frequently targeted perceived scientific consensus because raising awareness of scientific agreement can trigger a chain reaction: it increases belief that climate change is happening, human-caused and concerning, which in turn boosts support for public climate action (van der Linden et al., 2015, 2019). Despite this Gateway Belief Model being developed in relation to climate change, communicating scientific consensus has also proven effective in relation to other issues. For example, a recent study demonstrated that emphasising doctors' consensus on COVID-19 vaccines over a nine-month period led to a steady increase in vaccine uptake (Bartoš et al., 2022). Although there is a serious lack of field studies in this domain, these findings suggest that misinformation interventions targeting gateway beliefs may be successful in fostering actual shifts in behaviour. However, achieving the desired behavioural change in a sustained manner has proven extremely difficult, and unlikely to be achieved at scale with simple, one-off interventions[3].

2.2 Challenge #2: Testing Effects

Both human memory and attention are fallible, and one-off interventions are unlikely to yield sustained impact over time if no "booster shots" are administered. Maertens et al. (2024b) investigated effect decay over time of three "inoculation" intervention types (a video, a piece of text, and a game), and found that decay can set in within several days or sometimes weeks, but depends on the complexity of the desired outcome (e.g., remembering a single fact versus learning a more complex skill such as spotting a type of manipulation strategy).

Concerningly, however, testing effects are likely to interfere with the longevity of learning-based interventions. Capewell et al. (2024) and Maertens et al. (2024b) both found that various kinds of previously validated misinformation interventions (videos and games) are subject to rapid effect decay (with any effects dissipating in as little as 48 hours) if no immediate post-test (in the form of an item evaluation task) is administered. Capewell et al. (2024) also provide

[2] We note that Aird et al. (2018) showed that people's voting intentions did change when the number of false statements attributed to a politician vastly outweighed the number of truthful ones.
[3] It is worth noting that the ultimate goal of learning-based interventions is not (always) behavioural change, and so one could argue that any observed behavioural effects are interesting "side effects" as opposed to a core component of the intervention's effectiveness (Roozenbeek & van der Linden, 2024).

evidence for a causal mechanism: participants forget the relevant lessons from the intervention quickly if they do not rehearse what they learned immediately afterwards[4]. Theoretically, this means that many misinformation intervention studies may have overestimated the strength and longevity of the observed effects, as almost all of them administer item evaluation tasks immediately post-intervention even in longitudinal study designs. Practically, this means that especially interventions that do not easily incorporate practice (e.g., videos or text-based interventions) are likely less effective over time than earlier research indicated (e.g., Basol et al., 2021; Maertens et al., 2021).

With respect to nudges and other sharing-based interventions, Sasaki et al. (2021) found evidence for a reduction in the efficacy of nudge interventions with repeated exposure; however, few if any studies within the field of misinformation research explicitly test the possibility that nudges (or prompts) become less effective the more people see them[5]. In light of ongoing discussions about intervention "scalability" (Kozyreva et al., 2024; Roozenbeek et al., 2022a), it is possible that scalability only goes so far, and that the most scalable interventions may have fleeting real-world impact over time. This highlights the importance of conducting rigorous field research and longitudinal studies, and presents an impetus to move away from relying too much on (online) survey research when deciding "what works".

2.3 Challenge #3: Modest Impact

Individual-level intervention studies rely on the assumption that a substantial proportion of people who interact with them (e.g., in social media environments) do so as intended: they read the entire text, play the whole game to the end, or are duly reminded of the concept of accuracy. In the case of nudges, we assume that people are "nudgeable" (de Ridder et al., 2021) in that they not only engage in unwanted content sharing but are also amenable to intervention, which requires that their sharing behaviour must not be (entirely) habitual or deliberate (Ceylan et al., 2023). Brashier (2024) argues that "exposure prevalence" must therefore be taken into account before declaring that a misinformation intervention is a success, and that misinformation researchers ought to invest in designing interventions for those who need them the most.

An additional problem here is that, as all interventions are effective for a limited time period (Maertens et al., 2024b), we assume that a sufficient number of people not only encounter the type of unwanted content that the intervention tackles (e.g., "fake news" or a conspiracy theory) within this time frame, but *would have interacted with this content in an unwanted manner had it not been for the intervention*. After all, an intervention is ineffective for individuals who 1) engage in unwanted sharing or consumption behaviour, *and* 2) are potentially amenable to intervention, *and* 3) see the intervention, *and* 4) engage with the intervention as intended, *and* 5) encounter unwanted content relevant to the intervention before effect decay sets in, *but* 6) would have ignored this content anyway even without intervention. None of these nuances can be accurately explored in panel/survey studies. Put differently, the number of people for whom misinformation interventions are *practically* useful (and not just in theory) is likely to be lower than we think, and effect sizes obtained in survey studies are likely to be an overestimation of their actual effectiveness. We have visualised this problem in

[4] Capewell et al. (2024) argue that the item evaluation task itself can serve as sufficient practice for increasing longevity, but implementing such tasks as part of an intervention is impractical in real-world environments. Leder et al. (2024) found preliminary evidence that including a short feedback exercise at the end of a (gamified) intervention can boost its longevity.

[5] Roozenbeek et al. (2021) found preliminary evidence for rapid decay, i.e., a time span of several seconds, for a sharing-based "accuracy nudge" intervention, but more research is needed to test the robustness of this finding.

Figure 1.

To give an example of the above, Lin and colleagues (2024) report that content-neutral prompts reminding social media users to consider the accuracy of information significantly reduce the sharing of misinformation on X (formerly Twitter) and Facebook. However, this finding is limited: on Facebook[6], the interventions yielded only a very small effect when looking at all individuals who were targeted with the prompts in the first hour after being exposed to an accuracy prompt (i.e., they interacted with the intervention while it was 'active'). A stronger effect was detectable only for individuals who had shared misinformation in the week before the intervention (i.e., individuals who encounter unwanted content and have an imperfect ability to detect it as such, and/or share it out of ignorance or by accident), and also only in the first hour after intervention exposure. This study thus highlights the conditions required for misinformation interventions to successfully treat a comparatively small group of individuals, which underscores the need for complementary system-level interventions that have the potential to reach a larger proportion of the public (see Chater & Loewenstein, 2023).

2.4 Challenge #4: Reliance on item evaluation tasks

As mentioned above, item evaluation tasks take primacy in intervention efficacy evaluation. To be sure, this is preferable to not conducting randomised efficacy trials at all, as is too often the case when evaluating (for instance) educational or economic programmes in developing countries (Banerjee et al., 2010, 2016). However, it is also the case that other measures of efficacy are often ignored altogether, or play second fiddle to item evaluation task performance (Leder et al., 2024). But if a social media user ignores an intervention when they come across it online, it is practically useless, no matter the improvement in veracity discernment it yields in survey studies (which, again, are conducted under deliberately optimised conditions). However, with some exceptions (e.g., Davies et al., 2024), intervention *attractiveness* is almost never tested explicitly, nor is (potential) user uptake or retention (Johnson & Madsen, 2024). This is especially important in light of the above discussion about effect decay: one-off interactions are unlikely to yield impressive results, and so people must somehow be enticed to take "booster shots" (Maertens et al., 2024) or be "re-nudged" (Sasaki et al., 2021), which may be possible but is almost certainly subject to diminishing returns. The fact that these elements are rarely considered in intervention efficacy testing betrays a knowledge gap that is critical to overcome. Put simply, we argue that questions around what makes people amenable to interventions and how relevant audiences may be enticed to take part must start taking up space in misinformation research[7].

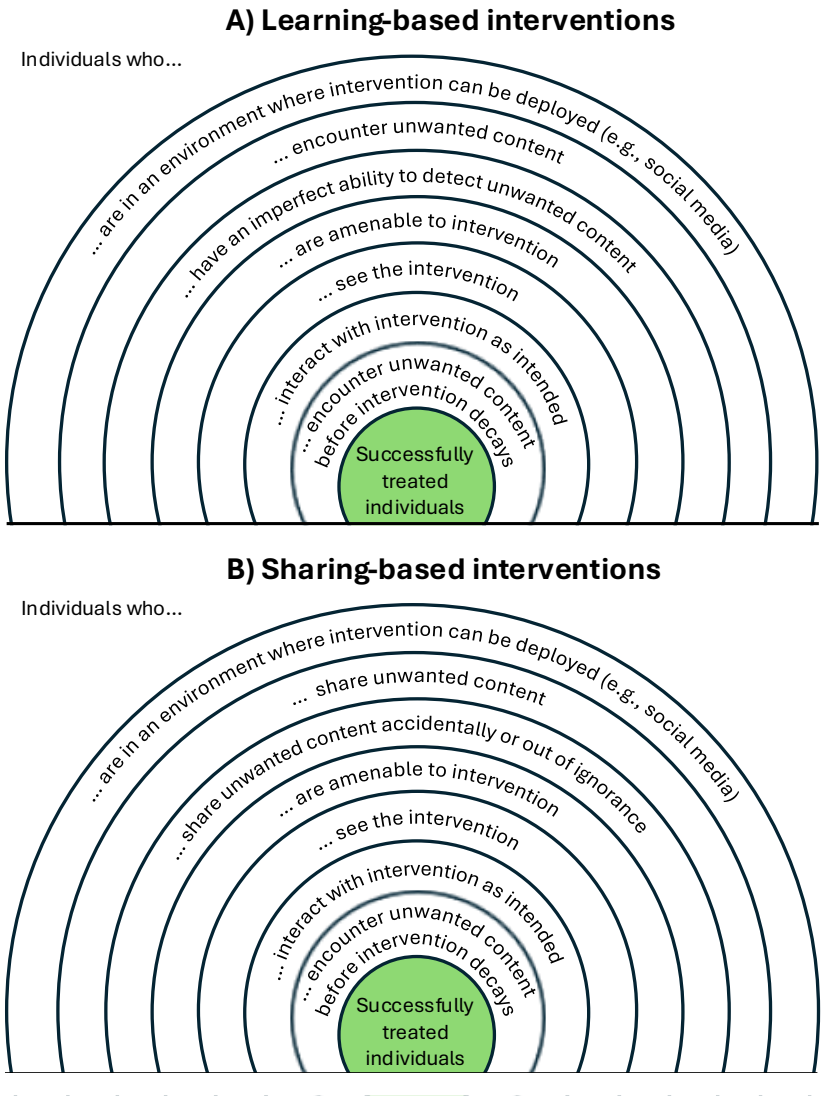
2.5 Challenge #5: "One Size Should Fit All"

Numerous researchers have attempted to transplant established interventions (shown to be effective in, say, the United States) onto non-Western contexts, with limited success: with some exceptions (e.g., Offer-Westort et al., 2024; Hopkins et al., 2023), replicability in the Global South remains low (Arechar et al., 2022; Guess et al., 2020; Harjani et al., 2023). However, there are nuances here that should not be ignored. For instance, Guess et al. (2020) found that media literacy tips were effective in India for a sample of highly educated individuals but not for a sample of individuals from a mostly rural area of the country; this may mean that the heart of the problem is not necessarily conducting research across borders, but rather that working with "hard to reach" populations continues to pose a substantial challenge.

[6] The authors report no effect decay over an 8-day period for their Twitter study.
[7] Here we follow Leder et al. (2024), who also "urge[d] the field to look beyond item rating task performance as a narrow measure of intervention efficacy".

Figure 1

*A chart showing which individuals are impacted by a hypothetical learning-based ("boosts"; **Panel A**) or sharing-based intervention ("nudges"; **Panel B**). The former intervention is aimed at boosting people's ability to recognise unwanted content (such as misinformation), the latter at preventing unwanted content sharing. Concentric half circles indicate diminishing potential group size. "Successfully treated individuals" are those whom the intervention successfully taught or boosted the target skill or competence (Panel A), or who were prevented from sharing unwanted content as a result of the intervention (Panel B). We make no claims as to the relative size of each concentric half circle, only that each successive group of individuals must logically be smaller than the one above it; we know, for instance, that the percentage of social media users who share "fake news" is low (Allen et al., 2020; Guess et al., 2019), and so it may be the case that only a fraction of social media users actually share unwanted content, and therefore that the second largest concentric circle in Panel B should be much smaller than the largest one, had they been scaled to relative size.*



As Badrinathan and Chauchard (2023) argue, the observed low replicability in Global South contexts should be no surprise: Western researchers operating in the Global South are often ill-equipped to translate both interventions and efficacy testing methods to local contexts. For example, Harjani et al. (2023) found that study participants from rural parts of India found their game (aimed at countering misinformation on WhatsApp) confusing, boring, and not relevant to their own lives. More broadly, Brashier (2024) argues that researchers should consider the preferences of the audiences they create the interventions for (not just Global South residents but also, for example, Conservatives in the United States). Here, too, many researchers assume that what works in one place ought to work (nearly) everywhere; after all, interventions are often grounded in psychological theories that should hold true for human beings in general. However, without conducting robust research into what works for whom and why, as well as increasing collaboration (and co-creation) with local researchers and implementation partners, limited cross-sectional efficacy is likely to continue.

This is not to say that successful examples of intervention studies from the Global South do not exist (Offer-Westort et al., 2024). For instance, Hopkins et al. (2023) tested the efficacy of an "inoculation" game about vaccine misinformation in Kenya, Uganda, and Rwanda, and found that the game generally improved discernment between factual information about vaccines and misinformation. This intervention, the *Cranky Uncle Vaccine* game, was developed as part of a lengthy co-designing process involving workshops with implementation partners and target audiences. Here, workshop participants provided iterative feedback about character design, gameplay experience, and content, thus helping to make the game more entertaining and relevant to potential players. This process contrasts with Harjani et al. (2023), who report that a similar gamified inoculation intervention (in their case about misinformation on WhatsApp) was ineffective at both engaging target audiences and effecting improved item evaluation task performance. It is possible that the game tested by Harjani et al. (2023) was not sufficiently tailored to local audiences due to the absence of a co-designing process. In other words, involving target audiences from the start, i.e., from the beginning of the intervention design process all the way to its implementation in the real world, is likely to benefit both an intervention's efficacy as well as its effectiveness.

2.6 Challenge #6: Unintended Consequences

Implementing misinformation interventions may also produce unexpected effects that go beyond merely failing to achieve the desired improvement in outcome measure performance. Following Nyhan and Reifler's (2010) series of experiments showing that misinformation interventions can accidentally strengthen some people's belief in targeted misconceptions, there has been concern about "backfire effects" in misinformation interventions research. Subsequent studies have called into question the prevalence and generalisability of such effects, demonstrating that people tend to heed corrective information even when it challenges their predispositions and that study design considerations are strongly linked to "backfire" effects being observed at all (Leder et al., 2024; Swire-Thompson et al., 2020, 2022; Wood & Porter, 2019).

However, it may be possible for misinformation interventions to have unintended effects beyond strengthening false beliefs, especially when implemented in contexts with which researchers lack familiarity. For example, scholars have commented that while falsehoods circulated within Western contexts are typically portrayed as isolated occurrences attributed to foreign influences or radical political entities, disinformation in non-Western societies is often depicted as stemming from undemocratic practices (Albuquerque, 2021). This suggests that if misinformation interventions – which

may be influenced by such underlying biases – are applied without adaptation in environments with political systems or cultural norms different from those that the designers of interventions are familiar with, they could be ineffective (Brashier, 2024). In extreme cases, interventions may even trigger reactance, expressed (for example) as heightened distrust against whoever is behind the intervention or increased and potentially unwarranted concern about the topic that the intervention is seeking to tackle. Johnson and Madsen (2024) report that people's self-reported likelihood of engaging with an "inoculation" intervention depends strongly on who developed it (which they call "inoculation hesitancy"); interestingly, even high-trust sources (e.g., Harvard University) did not outperform interventions that had no source at all, indicating that, in practice, people's (lack of) trust in who is behind the intervention (e.g., a tech company) may matter a great deal for maximising its effectiveness. More generally, item evaluation tasks are unable to pick up on the nuances of the effects conferred by misinformation interventions and cannot assess *why* an intervention might be (in)effective. The extent to which this and related phenomena might pose a problem for interventions when implemented in the real world is at present unknown, as the field is currently just beginning to take cross-cultural and cross-sectional research seriously (Brashier, 2024).

3. RECOMMENDATIONS AND CONCLUSION

Over the last decade, our understanding of how best to reduce the problem of online misinformation has flourished, and the interventions developed in this pursuit have become ever more sophisticated (Fazio et al., 2024). However, as we aim to further the reach, scalability, and durability of these interventions, it is important to acknowledge the limitations and assumptions underlying much of the current misinformation intervention literature (Brashier, 2024). At present, the status quo involves using item evaluation tasks to assess intervention efficacy with improved veracity or sharing discernment, presumed to translate to reduced belief in or sharing of misinformation "in the wild".

We raise six challenges for the research field. The first three relate to the fact that while misinformation interventions have at times produced impressive results on item evaluation tasks (thus demonstrating efficacy), these effects are likely to be inflated compared to the real-world impacts of many types of interventions. This is due to the artificially optimised nature of survey environments, the fallibility of human memory, the presence of testing effects, the practical difficulties associated with changing actual (as opposed to self-reported) behaviour and implementing intervention "booster shots", and the challenge of reaching those individuals who can be "successfully treated" against misinformation. The fourth challenge highlights how the field has so far regarded factors beyond item evaluation task performance as secondary, despite their important role in determining the practical effectiveness of interventions. These include intervention attractiveness, user uptake and retention, as well as familiarity with relevant cultural, socio-economic, and political contexts. Finally, interventions and the methods used to test their efficacy work less well outside of Western countries, and understanding and countering misinformation in the Global South (and more generally how to tailor interventions to specific audiences) remains understudied, which, in turn, may increase the risk of misinformation interventions having unintended consequences.

We propose several practical pathways for addressing these challenges. First, researchers and funders may consider investing in rigorous field research and assessing intervention *effectiveness* (rather than only *efficacy*). This may be achieved using a set of Key Performance Indicators, which could be estimated from data collected during lab studies and supplemented

Table 1

Challenges and recommendations for advancing the field of misinformation intervention studies.

Challenge #	Explanation	Recommendations
1: Prioritising efficacy over effectiveness	• There is a lack of (ecologically valid) field studies, and of those that exist, many yield either weak findings or null effects (particularly over time). • Successful lab studies are ubiquitous, but we lack knowledge on how lab results translate to real-world change.	• Consider what may make an intervention effective rather than merely efficacious. • Invest in field studies (e.g., social media interventions, school programmes). • Increase access to data (e.g., social media behavioural data). • Conduct more "many labs" and "many panels" studies.
2: Testing effects	• Item evaluation tasks administered immediately post-intervention substantially strengthen the intervention effect over time, as doing so helps people remember the intervention better. • Active rehearsal appears to play a key role in longevity, which impedes scalability.	• More research on the role of testing effects in intervention efficacy and longevity. • Explore innovative ways of incorporating rehearsal (e.g., gamification, quizzes).
3: Modest impact	• Individual-level interventions are likely to practically benefit only a small number of people (both in terms of boosting competencies and changing behaviour, see Figure 1). • Lab studies may overestimate intervention effect sizes due to optimised conditions.	• Tailor interventions to improve uptake by target audiences. • Invest in field studies; incorporate "system-level" and "individual-level" interventions. • Address both individuals and risk factors in their environment; engage with "high-impact" individuals ("super-spreaders" and role models).

	• Complexity of real-world environments may offset lab-observed effects.	• Explore ways of reaching critical mass for norm change.
4: Reliance on item evaluation tasks	• Most intervention studies use item evaluation task performance as the only or primary outcome measure of interest, and neglect user uptake, interest, attractiveness, reactance, and so on.	• Include additional outcome measures in lab studies. • Give more weight to measures other than item tasks, such as attractiveness and uptake potential.
5: "One size must fit all"	• The field continues to focus primarily on the US and a few other Western countries. • The few studies conducted in the Global South very often yield null findings. • Established testing methods may not be suitable for different contexts. • Interventions are often generalistic and/or assumed to work for everyone, and developing interventions for specific audiences and contexts remains challenging.	• Invest in "hard to reach" populations and geographical and sample diversity. • Conduct qualitative research into *why* interventions are (in)effective. • Invest in more ecologically valid testing methods. • Co-design and test interventions with local partners and target audiences.
6: Unintended consequences	• Outside of well-known "backfire effects" (which research shows are a minor concern), some interventions may fail to achieve their objectives due to reactance, different political/social/religious norms, socio-economic inequalities, and so on.	• Invest in qualitative research (e.g., focus groups, semi-structured interviews) examining the conditions in which interventions may be (in)effective. • Co-design interventions with local partners and audiences.

with publicly available statistics. An example is the probability of sharing a misinformation article for an average adult given one hour of social media exposure on a given platform, pre- and post-intervention. For instance, Allen et al. (2024) use machine learning and lab experiments to estimate that preventing exposure to vaccine-sceptical content could have resulted in 3 million more Americans getting vaccinated. Second, it is time to elevate additional outcome measures alongside item evaluation task performance. A practical first step could be to collect data on how likely an intervention is to be engaged with by the target audience by asking about its attractiveness, ease of implementation or interaction, and so on (Brashier, 2024; Davies et al., 2024; Johnson & Madsen, 2024). Third, to counteract testing effects, we recommend exploring innovative ways of incorporating active rehearsal, for example through quizzes (Lees et al., 2023) or gamification (Leder et al., 2024). Fourth, we recommend tailoring interventions to target audiences and combining "system-level" and "individual-level" interventions (Roozenbeek et al., 2023). Fifth, it is imperative to develop a better understanding of geographical and audience diversity, for example by co-designing interventions with local partners and audiences (Hopkins et al., 2023; for a guide, see Glennerster & Takavarasha, 2014). Finally, we must begin to prioritise qualitative, cross-sectional research (alongside the more conventional quantitative tests of efficacy) to better understand how information (including misleading or factually incorrect information shared by self-interested actors) impacts beliefs, attitudes, and behaviours. We have summarised our recommendations in Table 1.

Overall, we call for developing a framework for assessing intervention effectiveness beyond item evaluation tasks and for a more comprehensive definition of what counts as "success". To give a few examples: during the COVID-19 pandemic, health concerns might have driven belief in ineffective preventative measures such as drinking bleach (Delirrad & Mohammadi, 2020; Mahdavi et al., 2022), and in such cases intervention success may include a reduction in such unwanted behaviours. However, during political events such as the US Capitol riot of January 6th, 2021, politically motivated reasoning may play a more prominent role (Haslam et al., 2023). Therefore, an optimally successful intervention may also involve reducing "myside bias" (Roozenbeek et al., 2022b) or effecting depolarisation, alongside reducing false beliefs (Currin et al., 2022).

Table 1 is intended as a starting point for discussion. Fortunately, there are examples of misinformation interventions that can be considered efficacious as well as effective. For instance, the Stanford Civic Online Reasoning (COR) curriculum (https://cor.inquirygroup.org/) is a free set of courses and lessons to help students navigate online information, which has been shown to be efficacious when it comes to training lateral reading and general "internet savvy" (McGrew, 2020; McGrew et al., 2019; Wineburg et al., 2022). The COR also includes lessons and assessments for educators, and has been used in classrooms around the world. The same can be said for the News Evaluator (https://nyhetsvarderaren.se/in-english/) – a media literacy project led by researchers at the University of Uppsala (Axelsson et al., 2021, 2024). In both examples, lab results have been complemented with robust field trials (e.g., in schools), demonstrating efficacy, and interventions have been implemented in educational settings as part of a larger set of digital literacy initiatives. While these interventions are difficult to scale to whole populations compared to one-off nudges or video ad campaigns, they do demonstrate the potential for misinformation interventions to have sustained positive impact.

4. CONFLICTS OF INTEREST

The authors report no competing interests.

5. ACKNOWLEDGEMENTS

The authors report funding from the IRIS coalition (UK government, #SCH-00001-3391), JITSUVAX (EU Horizon 2020, #964728), and the Bill and Melinda Gates Foundation (#OPP1144).

6. AUTHOR CONTRIBUTIONS

J.R., Y.K., and M.R. conceptualised the paper and each author contributed to the writing and editing.

REFERENCES

- Aird, M. J., Ecker, U. K. H., Swire, B., Berinsky, A. J., & Lewandowsky, S. (2018). Does truth matter to voters? The effects of correcting political misinformation in an Australian sample. *Royal Society Open Science, 5*(12), 180593. https://doi.org/10.1098/rsos.180593
- Albuquerque, A. de. (2021). The institutional basis of anglophone western centrality. *Media, Culture, & Society, 43*(1), 180-188. https://doi.org/10.1177/0163443720957893
- Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14). https://doi.org/10.1126/sciadv.aay3539
- Allen, J., Watts, D. J., & Rand, D. G. (2024). Quantifying the impact of misinformation and vaccine-skeptical content on Facebook. *Science, 384*(6699), eadk3451. https://doi.org/10.1126/science.adk3451
- Arechar, A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M. N., Zhang, Y., Pennycook, G., & Rand, D. (2022). Understanding and Reducing Online Misinformation Across 16 Countries on Six Continents. *PsyArxiv Preprints*. https://doi.org/10.31234/osf.io/a9frz
- Aslett, K., Guess, A. M., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, *8*(18). https://doi.org/10.1126/sciadv.abl3844
- Axelsson, C.-A. W., Nygren, T., Roozenbeek, J., & van der Linden, S. (2024). Bad News in the civics classroom: How serious gameplay fosters teenagers' ability to discern misinformation techniques. *Journal of Research on Technology in Education*, 1–27. https://doi.org/10.1080/15391523.2024.2338451
- Axelsson, C.-A. W., Guath, M., & Nygren, T. (2021). Learning How to Separate Fake from Real News: Scalable Digital Tutorials Promoting Students' Civic Online Reasoning. *Future Internet*, *13*(3), 60. https://doi.org/10.3390/fi13030060
- Badrinathan, S., & Chauchard, S. (2023). Researching and Countering Misinformation in the Global South. *Current Opinion in Psychology*, 101733. https://doi.org/10.1016/j.copsyc.2023.101733
- Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. *American Economic Journal: Economic Policy*, *2*(1), 1–30. https://doi.org/10.1257/pol.2.1.1
- Banerjee, A. V., Duflo Esther, & Kremer, M. (2016). The Influence of Randomized Controlled Trials on Development Economics Research and on Development Policy. *"The State of Economics, The State of the World" Conference*. https://www.povertyactionlab.org/sites/default/files/research-paper/the-influence-of-rcts-on-developmental-economics-research-and-development-policy.pdf
- Bartoš, V., Bauer, M., Cahlíková, J., & Chytilová, J. (2022). Communicating doctors' consensus persistently increases COVID-19 vaccinations. *Nature, 606,* 542-549. https://doi.org/10.1038/s41586-022-04805-y
- Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data and Society*, *8*(1). https://doi.org/10.1177/20539517211013868
- Bauer, M. S., Damschroder, L., Hagedorn, H.,

Smith, J., & Kilbourne, A. M. (2015). An introduction to implementation science for the non-specialist. *BMC Psychology*, *3*(1), 32. https://doi.org/10.1186/s40359-015-0089-9

- Bauer, M. S., & Kirchner, J. (2020). Implementation science: What is it and why should I care? *Psychiatry Research*, *283*, 112376. https://doi.org/10.1016/j.psychres.2019.04.025
- Brauer, M. (2024). Stuck on Intergroup Attitudes: The Need to Shift Gears to Change Intergroup Behaviors. *Perspectives on Psychological Science, 19*(1), 280–294. https://doi.org/10.1177/17456916231185775
- 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
- Capewell, G., Maertens, R., Remshard, M., van der Linden, S., Compton, J., Lewandosky, S., & Roozenbeek, J. (2024). Misinformation interventions decay rapidly without an immediate post-test. *Journal of Applied Social Psychology*. https://doi.org/10.1111/jasp.13049
- Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), e2216614120. https://doi.org/10.1073/pnas.2216614120
- Chater, N., & Loewenstein, G. (2023). The i-Frame and the s-Frame: How Focusing on Individual-Level Solutions Has Led Behavioral Public Policy Astray. *Behavioral and Brain Sciences*, 46:e147. https://doi.org/10.1017/S0140525X22002023
- Currin, C. B., Vera, S. V., & Khaledi-Nasab, A. (2022). Depolarization of echo chambers by random dynamical nudge. *Scientific Reports*, *12*(1), 9234. https://doi.org/10.1038/s41598-022-12494-w
- Davies, B., Turner, M., & Udell, J. (2024). It helps to be funny or compassionate: An exploration of user experiences and evaluation of social media micro-intervention designs for protecting body image. *Computers in Human Behavior*, *150*, 107999. https://doi.org/10.1016/j.chb.2023.107999
- de Ridder, D., Kroese, F., & van Gestel, L. (2021). Nudgeability: Mapping Conditions of Susceptibility to Nudge Influence. *Perspectives on Psychological Science*, *17*(2), 346–359. https://doi.org/10.1177/1745691621995183
- Delirrad, M., & Mohammadi, A. B. (2020). New Methanol Poisoning Outbreaks in Iran Following COVID-19 Pandemic. *Alcohol and Alcoholism*, *55*(4), 347–348. https://doi.org/10.1093/alcalc/agaa036
- DellaVigna, S., & Linos, E. (2022). RCTs to Scale: Comprehensive Evidence from Two Nudge Units. *Econometrica*, *90*(1), 81–116. https://doi.org/10.3982/ECTA18709
- Done, A., Voss, C., & Rytter, N. G. (2011). Best practice interventions: Short-term impact and long-term outcomes. *Journal of Operations Management, 29*(5), 500–513. https://doi.org/10.1016/j.jom.2010.11.007
- Eccles, M. P., & Mittman, B. S. (2006). Welcome to Implementation Science. *Implementation Science*, *1*(1), 1. https://doi.org/10.1186/1748-5908-1-1
- Ecker, U. K. H., Roozenbeek, J., van der Linden, S., Tay, L. Q., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Misinformation poses a bigger threat to democracy than you might think. *Nature*, *630*(8015), 29–32. https://doi.org/10.1038/d41586-024-01587-3
- Fazio, L., Rand, D. G., Lewandowsky, S., Susmann, M., Berinsky, A. J., Guess, A. M., ... & Swire-Thompson, B. (2024). Combating misinformation: A megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/uyjha
- Glennerster, R., & Takavarasha, K. (2014). Running randomized evaluations: A practical guide. Princeton University Press.
- Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A.,

Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., … Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, *381*(6656), 398–404. https://doi.org/10.1126/science.abp9364

- Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (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. https://doi.org/10.1073/pnas.1920498117
- Guess, A. M., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1). https://doi.org/10.1126/sciadv.aau4586
- Hagger, M. S., & Weed, M. (2019). DEBATE: Do interventions based on behavioral theory work in the real world? *International Journal of Behavioral Nutrition and Physical Activity, 16*(1), 36. https://doi.org/10.1186/s12966-019-0795-4
- Harjani, T., Basol, M., Roozenbee, J., & van der Linden, S. (2023). Gamified inoculation against misinformation in India: a randomised control trial. *Journal of Trial and Error*. https://doi.org/10.36850/e12
- Haslam, S. A., Reicher, S. D., Selvanathan, H. P., Gaffney, A. M., Steffens, N. K., Packer, D., Van Bavel, J. J., Ntontis, E., Neville, F., Vestergren, S., Jurstakova, K., & Platow, M. J. (2023). Examining the role of Donald Trump and his supporters in the 2021 assault on the U.S. Capitol: A dual-agency model of identity leadership and engaged followership. *The Leadership Quarterly*, *34*(2), 101622. https://doi.org/10.1016/j.leaqua.2022.101622
- Hobby, J., Crowley, J., Barnes, K., Mitchell, L., Parkinson, J., & Ball, L. (2022). Effectiveness of interventions to improve health behaviours of health professionals: a systematic review. *BMJ Open, 12*(9), e058955. https://doi.org/10.1136/bmjopen-2021-058955
- Hopkins, K. L., Lepage, C., Cook, W., Thomson, A., Abeyesekera, S., Knobler, S., Boehman, N., Thompson, B., Waiswa, P., Ssanyu, J. N., Kabwijamu, L., Wamalwa, B., Aura, C., Rukundo, J. C., & Cook, J. (2023). Co-Designing a Mobile-Based Game to Improve Misinformation Resistance and Vaccine Knowledge in Uganda, Kenya, and Rwanda. *Journal of Health Communication*, *28*(sup2), 49–60. https://doi.org/10.1080/10810730.2023.2231377
- Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Maertens, R., Panizza, F., Pennycook, G., Rand, D. J., Rathje, S., Reifler, J., … Wineburg, S. (2024). Toolbox of Interventions Against Online Misinformation and Manipulation. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-01881-0
- Johnson, A., & Madsen, J. K. (2024). Inoculation hesitancy: an exploration of challenges in scaling inoculation theory. *Royal Society Open Science*, *11*(6). https://doi.org/10.1098/rsos.231711
- Leder, J., Schellinger, L. V., Maertens, R., Chryst, B., van der Linden, S., & Roozenbeek, J. (2024). Feedback Exercises Boost Misinformation Discernment for Gamified Interventions. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001603
- Lees, J., Banas, J. A., Linvill, D., Meirick, P. C., & Warren, P. (2023). The Spot the Troll Quiz Game Increases Accuracy in Discerning Between Real and Inauthentic Social Media Accounts. *PNAS Nexus*. https://doi.org/10.31219/osf.io/xu6mh
- Lin, H., Garro, H., Wernerfelt, N., Shore, J. C., Hughes, A., Deisenroth, D., … Rand, D. G. (2024). Reducing misinformation sharing at scale using digital accuracy prompt ads. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/u8anb

- Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X.-D. (2023). Psychological Inoculation for Credibility Assessment, Sharing Intention, and Discernment of Misinformation: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *25*, e49255. https://doi.org/10.2196/49255
- Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., He, J., & van der Linden, S. (2024a). The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods, 56*, 1863–1899. https://doi.org/10.3758/s13428-023-02124-2
- Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315
- Maertens, R., Roozenbeek, J., Simons, J., Lewandowsky, S., Maturo, V., Goldberg, B., Xu, R., & van der Linden, S. (2024b). Psychological Booster Shots Targeting Memory Increase Long-Term Resistance Against Misinformation. *PsyArxiv Preprints*.
- Mahdavi, S. A., Zamani, N., McDonald, R., Akhgari, M., Kolahi, A.-A., Gheshlaghi, F., Ostadi, A., Dehghan, A., Moshiri, M., Rahbar-Taramsari, M., Delirrad, M., Mohtasham, N., Afzali, S., Ebrahimi, S., Ziaeefar, P., Khosravi, N., Kazemifar, A. M., Ghadirzadeh, M., Farajidana, H., … Hassanian-Moghaddam, H. (2022). A cross-sectional multicenter linkage study of hospital admissions and mortality due to methanol poisoning in Iranian adults during the COVID-19 pandemic. *Scientific Reports*, *12*(1), 9741. https://doi.org/10.1038/s41598-022-14007-1
- McGrew, S. (2020). Learning to evaluate: An intervention in civic online reasoning. *Computers & Education*, *145*, 103711. https://doi.org/10.1016/j.compedu.2019.103711
- McGrew, S., Smith, M., Breakstone, J., Ortega, T., & Wineburg, S. (2019). Improving university students' web savvy: An intervention study. *British Journal of Educational Psychology*, *89*(3), 485–500. https://doi.org/10.1111/bjep.12279
- McPhedran, R., Ratajczak, M., Mawby, M., King, E., Yang, Y., & Gold, N. (2023). Psychological inoculation protects against the social media infodemic. *Scientific Reports, 13*(1), 5780. https://doi.org/10.1038/s41598-023-32962-1
- Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303-330. https://doi.org/10.1007/s11109-010-9112-2
- Offer-Westort, M., Rosenzweig, L. R., & Athey, S. (2024). Battling the coronavirus 'infodemic' among social media users in Kenya and Nigeria. *Nature Human Behaviour*, *8*(5), 823–834. https://doi.org/10.1038/s41562-023-01810-7
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590–595. https://doi.org/10.1038/s41586-021-03344-2
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011
- Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*, 2333. https://doi.org/10.1038/s41467-022-30073-5
- Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering Misinformation: Evidence,

Knowledge Gaps, and Implications of Current Interventions. *European Psychologist*, *28*(3), 189–205. https://doi.org/10.1027/1016-9040/a000492

- Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy nudges? A pre-registered direct replication of Pennycook et al. (2020). *Psychological Science*, *32*(7), 1–10. https://doi.org/10.1177/09567976211024535
- Roozenbeek, J., Maertens, R., Herzog, S., Geers, M., Kurvers, R., Sultan, M., & van der Linden, S. (2022b). Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, *17*(3), 547–573. https://doi.org/10.1017/S1930297500003570
- Roozenbeek, J., & van der Linden, S. (2024). *The Psychology of Misinformation*. Cambridge University Press.
- Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022a). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34). https://doi.org/10.1126/sciadv.abo6254
- Sasaki, S., Kurokawa, H., & Ohtake, F. (2021). Effective but fragile? Responses to repeated nudge-based messages for preventing the spread of COVID-19 infection. *The Japanese Economic Review*, *62*, 371–408. https://doi.org/10.1007/s42973-021-00076-w
- Singal, A. G., Higgins, P. D. R., & Waljee, A. K. (2014). A Primer on Effectiveness and Efficacy Trials. *Clinical and Translational Gastroenterology, 5*(1), e45. https://doi.org/10.1038/ctg.2013.13
- Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: Comprehending the Trump phenomenon. Royal Society Open Science, 4(3), 160802. http://doi.org/10.1098/rsos.160802
- Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognitions, 9*(3), 286-299. https://doi.org/10.1016/j.jarmac.2020.06.006
- Swire-Thompson, B., Miklaucic, N., Wihbey, J., Lazer, D., & DeGutis, J. (2022). Backfire effects after correcting misinformation are strongly associated with reliability. *Journal of Experimental Psychology: General*, *151*(7), 1655–1665. https://doi.org/10.1037/xge0001131
- van der Linden, S. L., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PloS ONE, 10*(2), e0118489. https://doi.org/10.1371/journal.pone.0118489
- van der Linden, S., Leiserowitz, A., & Maibach, E. (2019). The gateway belief model: A large-scale replication. Journal of Environmental Psychology, 62, 49-58. https://doi.org/10.1016/j.jenvp.2019.01.009
- Wineburg, S., Breakstone, J., McGrew, S., Smith, M., & Ortega, T. (2022). Lateral reading on the open Internet: A district-wide field study in high school government classes. *Journal of Educational Psychology*, *114*(5), 893–909. https://doi.org/10.1037/edu0000740
- Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior, 41*, 135-163. https://doi.org/10.1007/s11109-018-9443-y