INTERVENTIONS TO COUNTER MISINFORMATION: LESSONS FROM THE GLOBAL NORTH AND APPLICATIONS TO THE GLOBAL SOUTH

JULY 2023

Prepared under Contract No.: GS-10F-0033M / 7200AA18M00016, Tasking N067

DRG LEARNING, EVALUATION, AND RESEARCH ACTIVITY II

INTERVENTIONS TO COUNTER MISINFORMATION: LESSONS FROM THE GLOBAL NORTH AND APPLICATIONS TO THE GLOBAL SOUTH

JULY 2023

Prepared under Contract No.: GS-10F-0033M /7200AA18M00016, Tasking N067

Submitted to:

Matthew Baker, USAID COR

Submitted by:

Robert A. Blair[1], Jessica Gottlieb[2], Brendan Nyhan[3], Laura Paler[4], Pablo Argote[5], Charlene J. Stainfield[6]

Contractor:

NORC at the University of Chicago
4350 East West Highway, 8th Floor
Bethesda, MD 20814
Attention: Matthew Parry
Tel: 301- 634-5444; E-mail: Parry-Matthew@norc.org

DISCLAIMER

The authors' views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

[1] Arkadij Eisler Goldman Sachs Associate Professor, Department of Political Science and Watson Institute for International and Public Affairs, Brown University
[2] Associate Professor, Hobby School of Public Affairs, University of Houston
[3] James O. Freedman Presidential Professor, Department of Government, Dartmouth College
[4] Associate Professor, Department of Government, School of Public Affairs, American University
[5] Post-Doctoral Scholar, Department of Political Science and International Relations, University of Southern California
[6] Ph.D. Candidate, Department of Political Science, Ohio State University

TABLE OF CONTENTS

1. Executive summary 3
2. Introduction 5
3. Scope and methodology 10
4. Findings 13
4.1 Informational interventions 13
4.1.1 Inoculation 13
4.1.2 Debunking 19
4.1.3 Credibility labels and tags 25
4.1.4 Contextual labels and tags/provenance cues 28
4.2 Educational interventions 30
4.2.1 Media literacy 30
4.3 Socio-psychological interventions 36
4.3.1 Accuracy prompts 36
4.3.2 Friction 39
4.3.3 Social norm prompts 40
4.4 Institutional interventions 43
4.4.1 Platform alterations 43
4.4.2 Politician messaging 44
4.4.3 Journalistic interventions 46
5. Expert survey results 49
5.1 Quantitative analysis 49
5.2 Qualitative analysis 52
6. How context moderates effectiveness of interven- tions 55
7. Discussion and recommendations 56
7.1 Practical considerations for implementation 56
7.2 Promising areas for future research 57
7.2.1 Designing studies that allow for more direct comparisons between Global North and Global South contexts 58
7.2.2 Exploring whether some types of misinformation are easier to curb than others 58
7.2.3 Understanding the role of social identity in efforts to combat misinfor- mation 59
7.2.4 Testing whether interventions are more effective in combination with one another 59
7.2.5 Expanding the evidence base on understudied interventions 59
References 61

TABLES

Table 1: Summary of findings for interventions 8
Table 2: Matrix of research questions 10
Table 3: Summary of intervention types 12
Table 4: Description of media literacy interventions 31
Table 5: Qualitative assessments of interventions 58

FIGURES

Figure 1: Inoculation treatment from Vivion et al. (2022) (boxes added) 15
Figure 2: Screenshots from inoculation games (Maertens et al., 2021; Basol et al., 2021) 17
Figure 3: Debunking treatments from Porter et al. (2023) 23
Figure 4: Disputed tags from Mena (2020) 26
Figure 5: Context labels from Nassetta and Gross (2020) 29
Figure 6: Media literacy treatment in the U.S. from Qian et al. (2022) 33
Figure 7: Media literacy treatment in Nigeria from Apuke et al. (2023) 35
Figure 8: Accuracy prompt treatments from Epstein et al. (2021) 36
Figure 9: Friction intervention on Twitter from Sharevski et al. (2022) 40
Figure 10: Experimental manipulation from Clayton et al. (2019) 42
Figure 11: Artificial constituent letter from Diermeier (2023) (box added) 46
Figure 12: Journalistic intervention from Graves et al. (2016) (box added) 47
Figure 13: Categories of respondents in the expert survey 50
Figure 14: Mean allocations to each intervention type by respondent category 51
Figure 15: Extent to which interventions would be *less* effective in a developing country context compared to a developed one 53
Figure 16: Comparing expert evaluations with quantity of evidence 60

1. EXECUTIVE SUMMARY

The production and spread of misinformation can have harmful effects on democracy, social cohesion, trust in institutions, and public health outcomes. It is thus vitally important to identify effective strategies for countering misinformation. Though a robust literature examines how to combat misinformation in the Global North, substantially less attention has been paid to the strategies and interventions that might be most effective in the Global South. This gap in scientific knowledge is alarming, especially as misinformation proliferates across Global South countries. We should not assume that interventions which work in the Global North will necessarily be effective elsewhere.

To better understand how to counter misinformation in the Global South, this review synthesizes evidence from 176 intervention tests reported in 155 unique studies conducted in both the Global North and Global South. We focus on randomized control trials (RCTs) because our main goal is to evaluate the causal effects of strategies to counter misinformation, which requires a control group. We review evidence from RCTs on 11 leading interventions, which we classify into four broad families: (1) informational interventions (inoculation, debunking, credibility labels, and contextual labels); (2) educational interventions (media literacy); (3) socio-psychological interventions (accuracy prompts, friction, and social norms interventions); and (4) institutional interventions (platform alterations, politician messaging, and journalistic/media interventions).

Our review of the evidence yields several main conclusions:

There is indeed an acute gap in evidence on what works to counter misinformation in the Global North versus the Global South. Of the 155 studies included in this review, over 80% were conducted in one or more Global North countries. This severe imbalance in evidence quantity highlights the challenges of drawing conclusions about effective strategies for countering misinformation in the Global South.

There is also substantial variation in the strength of the evidence base across leading interventions. Some interventions have been relatively well studied — for example, 56 of 155 unique studies included in this review test the effectiveness of debunking. But other promising interventions like media literacy are relatively understudied. Institutional interventions, which arguably hold the greatest promise, have been studied the least, with no studies on politician messaging or journalist training having been conducted in Global South countries to date.

There is strong evidence that inoculation (prebunking) and debunking are generally effective at countering misinformation in both Global North and Global South countries. Together, these two informational interventions are the most widely studied among all interventions included in this review. Moreover, the majority of studies suggest that both interventions reduce belief in misinformation. Both inoculation and debunking also appear to be more effective than other informational interventions, like credibility and contextual labels. These effects are strongest in the short term, however, and typically measured only among people who are exposed to the debunking or inoculation (thus abstracting away from dissemination challenges in the real world). Moreover, effects vary by factors such as the type of correction (e.g., providing factual information versus explaining how a deception technique works) and the source of the correction.

Evidence for the effectiveness of media literacy interventions, which is one of the most popular strategies for combating misinformation, is mixed. Overall, relatively few studies included in this review have assessed the effectiveness of media literacy interventions. Nevertheless, several of these studies are especially high-quality, including eight studies conducted in five Global South countries. The majority of studies on media literacy in the Global North find no effect on misinformation beliefs. Evidence on effects in the Global South is mixed, with some evidence that media literacy works but only among populations with already high baseline literacy and education levels.

Interventions that alter social norms around misinformation have potential, although more evidence is needed. We define socio-psychological interventions as those that try to change beliefs or behaviors surrounding misinformation by affecting people's mindsets or invoking social identity without necessarily providing them with new information or skills to counter misinformation. The evidence we review shows that both accuracy prompts (which remind people of the importance of correct information) and friction interventions (which encourage people to slow down and consider the information they engage with) are effective at increasing discernment between true and false information. Our findings suggest that social norms interventions, which typically use messaging from in-group members to change perceived norms around sharing or believing misinformation, have even greater potential. There is compelling evidence from the Global North that social norms interventions are generally effective at reducing both misinformed beliefs and sharing intentions. While the Global South literature on these interventions is still small and results are mixed, this intervention represents a promising avenue for future research because of the relative strength of community ties.

Institutional interventions, which have perhaps the greatest potential to affect change, have been the least studied, with few Global North and no Global South studies to date on politician messaging or journalist training. Interventions that aim to alter the supply of misinformation by platforms, politicians, or journalists are often harder to implement and evaluate than those discussed above, which generally aim to reduce the demand for misinformation at the individual level. There is a dramatic need to improve the evidence base on these interventions.

The least evidence exists about the interventions experts think will work best.

To assess which interventions would be effective in the Global South, we complemented our literature review with a survey of 138 experts who work in both research and practice in areas related to misinformation and governance. Strikingly, experts express the greatest optimism about the interventions for which the least evidence exists. Specifically, the three most popular interventions among experts — media literacy, journalist training, and platform alterations — are among the least studied, with a total of 29 studies between them. Conversely, and perhaps most surprisingly, experts were least optimistic about informational and socio- psychological interventions, the interventions which have been studied most and have the strongest record of effectiveness.

Context matters when trying to assess whether findings travel from the Global North to the Global South or across Global South countries. Because of the lack of studies on countering misinformation in the Global South, it is tempting to generalize findings from the Global North to the Global South or between Global South countries. For instance, if evidence suggests that media literacy interventions work on educated populations in India, would we see similar results if the same intervention were conducted in countries in Southeast Asia or Sub- Saharan Africa? Alternatively, imagine that evidence showed journalist training interventions generally work in liberal democracies — should we also expect

them to work in more closed regimes? Answering questions like these is an important but inherently speculative exercise that can be made more rigorous by examining two relevant sets of considerations. First, what contextual factors at the individual level (e.g., educational attainment) and/or country level (e.g., regime type) likely matter most to the effectiveness of an intervention? Second, to what extent are different contexts similar on these key dimensions such that findings from one context might reasonably be expected to hold in similar cases? To facilitate more rigorous thinking about these difficult questions, we invite readers to use a database that we created to accompany this report. The database allows users to filter studies by key country and population-level characteristics to find the evidence that speaks most directly to the contexts and cases that they identify as most relevant.

2. INTRODUCTION

This evidence review addresses one of the overarching questions in the 2021– 2023 USAID Center for Democracy, Human Rights, and Governance (DRG) Learning Agenda: *What factors and dynamics foster—and build resilience to—the proliferation of disinformation, misinformation, and/or malinformation?*

Misinformation can be defined as factual claims about the world that are either strictly false or contradicted by high-quality evidence and expert opinion (see Vraga and Bode 2020 for a discussion of definitional challenges in this area). Related concepts include disinformation (intentionally false misinformation) and conspiracy theories (attributions of events to the secret actions of powerful people or organizations); popular related terms such as propaganda and "fake news" are often frequently invoked but typically imprecisely defined.

Academic research on misinformation and mechanisms to combat it in the U.S. or the Global North have proliferated wildly in recent years. Accordingly, numerous literature reviews have already been published (e.g., Nyhan, 2020, 2021; Pennycook and Rand, 2021; Ecker et al., 2022; Johansson et al., 2022; Kozyreva et al., 2022; Van Der Linden, 2022). By contrast, there is still only a nascent literature studying misinformation in the Global South, a term for developing countries in East Asia, Latin America, the Middle East and North Africa (MENA), South Asia, and sub- Saharan Africa (see Finance Center for South-South Cooperation 2023 for one list of qualifying countries). This emerging literature reveals a mixed picture of whether and how we can apply lessons from the Global North to the Global South. It also substantiates the need for a new literature review that focuses on these contextual differences and discusses how programming may need to be adapted to be most effective in the contexts in which USAID operates.

The need for research on misinformation in the Global South is acute; it has proven to be a serious challenge with sometimes devastating consequences. In Brazil, supporters and affiliates of then-candidate Jair Bolsonaro used WhatsApp to disseminate misinformation questioning the integrity of the 2018 presidential election, smearing rival candidates, and attacking the legitimacy of the media (Bandeira et al., 2019). In early 2021, a video implying that Ivoirian migrants were being attacked in neighboring Niger prompted acts of violence against Nigerians in Côte d'Ivoire's capital, Abidjan. It was later revealed that the video was actually two years old and depicted terrorist arrests in Nigeria (APA News, 2021). In India, the fact-checking organization *Alt News* has come under intense government scrutiny for attempting to debunk rumors related to child kidnapping gangs, the role of the country's Muslim population in spreading COVID-19, and other conspiracy theories, some of which have gained enough traction to become mainstream TV news stories (Raj, 2022). Facebook's parent company Meta is currently embroiled

in a lawsuit accusing the platform of amplifying hate speech and disinformation that contributed to the genocide against Myanmar's Rohingya minority group (Amnesty International, 2022); Facebook was similarly accused of facilitating the spread of misinformation that resulted in widespread ethnic violence in Ethiopia (Jackson et al., 2022).

Interventions designed to combat misinformation are likely to work differently across contexts, particularly when moving from the Global North to the Global South. First, people in the Global South use WhatsApp relatively more than Facebook or Twitter, making it more difficult to observe (much less control) the flow of misinformation. Second, lower average levels of literacy, including digital and media literacy, could simultaneously make people more vulnerable to misinformation and also potentially increase the effects of interventions to promote digital and media literacy.

Third, state weakness in many countries of the Global South leads to greater reliance on social institutions, such as ethnic or religious groups, whose leaders function as intermediaries between the state and its citizens. Again, this contextual difference has potentially countervailing implications. These institutions and intermediaries could be valuable partners in interventions to combat misinformation, but they may also help create a degraded information environment in which people are more skeptical and distrustful of all truth claims (Altay et al., 2023). Finally, in younger democracies, independent sources of media or trusted adjudicators of the truth may be rarer, potentially increasing the importance of source trust when debunking false claims.

The main goals of this review are thus twofold: (1) to synthesize and translate lessons about the distribution, reception, and correction of misinformation from the Global North to the Global South and (2) to review evidence from the Global South to draw conclusions about which interventions are most likely to be effective there. To accomplish these goals, we first draw on the existing literature to identify 11 leading interventions. We classify these into four main intervention categories based on the actors targeted by the intervention (i.e., consumers or producers of misinformation) and the nature of the intervention itself.

The first three categories of interventions primarily target individuals with the goal of making them less likely to believe or share misinformation. Of these, *informational interventions* like inoculation and debunking provide corrective information that aims to neutralize specific pieces of misinformation. In contrast, *educational interventions*, of which media literacy is the leading example, seek to provide individuals with a broader set of skills to make them less susceptible to misinformation. *Socio-psychological interventions* use priming or appeals to social identity and social costs to discourage the take-up and spread of misinformation. By contrast, interventions in the fourth category—*institutional interventions*—instead seek to change the behavior of the producers and distributors of misinformation, including platforms, politicians, and media professionals.

We assess the effectiveness of the 11 interventions we identified in these categories based on a review of 155 unique studies conducted in the Global North and, to some extent, the Global South. Our main findings are summarized in Table 1 and discussed in each section below. Key conclusions from Table 1 are summarized below:

- **Debunking and inoculation work**. These two informational interventions have the strongest evidence base — more than 70 unique studies — and are frequently effective at reducing false

beliefs. However, only 16 of the unique studies tested debunking and/or inoculation in the Global South.

- **Evidence on media literacy**, the leading educational intervention, is **mixed** in the Global North and Global South.
- **Social norm interventions appear most effective among the socio-psychological interventions.**
- **Institutional interventions have the greatest potential impact but the least evidence** from either the Global North or South.

Table 1: Summary of findings for interventions

Category	Intervention	Global North finding	Global South finding
Informational	Inoculation	Generally effective in the short term, although evidence is mixed regarding technique inoculation. Inoculation games may be less effective than first thought.	Tends to be most successful when implemented over long stretches of time in partnership with local news entities.
Informational	Debunking	Generally effective at reducing misinformed beliefs but effects may vary depending on credibility of or trust in the correction source.	More effective when delivered or endorsed by a trusted or in-group source; mixed evidence regarding effect on behavioral intentions.
Informational	Credibility labels/tags	Most effective when labels/tags provide justification for their presence, provide clear true or false ratings, and reference expert fact-checkers.	One study shows the potential for labels to decrease individual sharing of misinformation; more evidence needed.
Informational	Contextual labels/tags	Mixed effects on belief and sharing intentions; though effect sizes vary.	One study shows reduced perceived credibility misinformation; more evidence needed.
Educational	Media literacy	Limited evidence of effectiveness and substantial variation in durability.	Mixed evidence of effectiveness despite intensive nature of some interventions
Socio-psychological	Accuracy prompts	Generally effective at increasing discernment, though effect sizes vary	Sometimes effective in increasing discernment and reducing sharing, but effects are small.
Socio-psychological	Friction	May increase truth discernment and reduce sharing intentions for false news; more evidence needed.	No evidence; studies needed.
Institutional	Platform Alterations	Generally positive, though most evidence gathered from simulating platforms as opposed to actual platform change.	Experimental evidence needed.
Institutional	Politician messaging	Shows initial promise at reducing misinformation supply; more evidence needed.	No evidence; studies needed.
Institutional	Journalist training	Shows initial promise for combating misinformation; more evidence needed.	No evidence; studies needed.

We note, however, critical contextual points that must be kept in mind:

- Studies from the Global South are newer on average and thus more likely to be unpublished, potentially increasing the likelihood of reporting null results (given publication bias toward positive findings in the set of older studies that are publicly available). Differences in results may thus be attributable in some cases to differences in publication status rather than contextual differences between the Global North and Global South. (Further discussion of this point is provided in the scope and methodology section below.)

- Resource allocation decisions should keep in mind the implementation issues we discuss below, which are especially acute for public-facing interventions. Many are difficult to deliver at the right time (e.g., right before or after exposure) to the audience that is most vulnerable to misinformation. These constraints are rarely addressed directly in the designs of the studies we consider, which instead typically control both exposure and intervention delivery and can thus abstract away from these issues. (Further discussion of this point is provided in the discussion and recommendations section below.)

To facilitate deeper understanding of these results and how they vary by context, this evidence review is accompanied by a searchable database that can be accessed here[7]. The database allows users to filter the set of identified studies using key variables on the study context — regime type, degree of media freedom, GDP per capita, and internet penetration in the country-year in which the study was conducted — to find evidence that is most relevant to their needs. The full dataset is described below in Section 6.

We complement this evidence review with data from an original survey in which we asked experts to identify the most promising interventions for reducing misinformation in the Global South. We analyze patterns in the expert survey in Section 5, including which interventions were expected to be most and least promising on average, and the extent to which experts differed based on their role (researchers versus practitioners) and area of expertise (experts on the Global North versus South). Notably, we find the interventions that experts expect to be most effective in the Global South are also the interventions with the least existing evidence, which provides important guidance for future research expenditures. We also solicited qualitative feedback in open-ended questions about how context was likely to moderate intervention effectiveness. These findings, which are reported in Section 6, provide testable intuitions about potential moderators of the effects of misinformation interventions that can be evaluated in future research.

[7] Full link to the database: https://www.democratic-erosion.com/briefs/ misinformation-intervention-database/

3. SCOPE AND METHODOLOGY

Our methodological approach to the literature review is motivated by our twin goals of (1) synthesizing and translating lessons about the distribution, reception, and correction of misinformation from the Global North to the Global South and (2) reviewing the evidence from Global South contexts. First, we drew on literature reviews with evidence largely from the Global North to identify the most important misinformation interventions that have been tested (e.g., accuracy prompts, debunking, labels, inoculation, media literacy, frictions, social norms, and institutional interventions). For each intervention, we describe the theoretical justification underlying it; evaluate the existing evidence with particular attention to evidence from non-Western contexts when available; comment on the intervention's applicability to the Global South, especially when the only existing evidence is from the Global North; and discuss the feasibility of implementing the intervention, with attention given to what is known (and what remains unknown) regarding the scalability and duration of effects.

Table 2 describes our key research questions and data sources:

Table 2: Matrix of research questions

Question	Data source
What are the main types of interventions that have been tested to reduce misinformation?	**Existing literature reviews (Global North)**
What evidence exists about these interventions from the Global South?	**Literature review, network of experts**
Which interventions are most promising for application in the Global South?	**Literature review, experts on misinformation and governance in the Global South**

The first component of our data collection was a literature search that was intended to ensure we captured research from both the Global North and Global South. The process was iterative. To construct a comprehensive framework of intervention categories, we first read through meta-analyses and literature reviews of the misinformation literature and created a list of 11 interventions and mechanisms. We then categorized these interventions into four types of interventions: informational, educational, socio-psychological, and institutional.

Once the intervention framework was finalized, we identified search terms for each intervention that were used on Google Scholar and Elicit. For the Global North, the search terms were structured as follows, with backslashes indicating separate searches per term: "*[intervention name]* + misinformation / disinformation / malinformation / fake news / false news". We intentionally crafted specific search terms for the Global South that would yield as much evidence as possible: "*[intervention name]* + misinformation / disinformation / malinformation / fake news / false news + Global South / developing country(ies) / East Asia / Latin America / Middle East and North Africa / South Asia / sub-Saharan Africa". Searches for interventions that have been well-studied in the Global North included an added filter for year to focus on research conducted after recent meta-analyses and literature reviews (2017–2023

for debunking given Chan et al. 2017, 2020– 2023 for credibility labels and tags given Walter et al. 2020, and 2021–2023 for inoculation given Banas and Rains 2010 and Compton et al. 2021).

Importantly, this report is not a formal (quantitative) meta-analysis: our conclusions about the efficacy and broad applicability of each intervention are based on our own qualitative assessments of the literature, which take into account the proportion of studies suggesting positive, negative, and null effects of each intervention; the strength and durability of the estimated effects; and the quality and methodological rigor of the underlying studies. For example, if we find three studies suggesting that an intervention is effective and two suggesting that it is not, but if the two studies showing null or negative effects are much more rigorous than the three showing positive effects, we might conclude that the intervention is probably not effective, at least in the contexts where it has been evaluated.

Relatedly, the evidence presented here is limited to randomized controlled trials (RCTs), which are the simplest way to estimate causal effects. Because RCTs randomly assign units to treatment and control, fewer assumptions are needed to establish that the control group serves as a good counterfactual for the treatment group and thus that a difference in outcomes between treatment and control can be interpreted as causal. The team elected not to consider quasi-experimental studies, observational studies, or program evaluations that lacked random assignment to treatment or a clear control. In these study designs, the control group can differ systematically from the treatment group in a way that produces differences in outcomes that are not due to the treatment and that are hard to account for via control variables or other research design approaches. We focus on RCTs because it would be challenging or impossible to effectively evaluate the credibility of the stringent assumptions required to interpret non-experimental findings as causal within the scope of our review.

The RCTs we consider evaluate a number of different outcome measures. The most common are belief in false claims and intention to share false claims as expressed in surveys. However, the most compelling evidence comes from studies that demonstrate that the intervention in question improves people's ability to distinguish between true and false information (what the literature calls "discernment"). Failing to measure discernment can lead scholars and practitioners to falsely believe a treatment is reducing belief in (or sharing of) misinformation when it is actually causing people to distrust *all* information (as shown in a reanalysis of the effects of exposure to inoculation games by Modirrousta-Galian and Higham 2023). Other outcomes that are considered include policy attitudes, evaluations of people or groups, vote choice, and real-world behaviors such as public statements or actions on social media.

We also note that there are potentially important differences between the RCTs considered in our review from the Global South and the Global North. The studies in the Global South that we review here were conducted more recently on average than those conducted in the Global North. As a result, more of them are unpublished and thus potentially more likely to show null results due to the bias in the scientific peer review and publication process toward significant findings. For these reasons, we urge caution in interpreting null results from the Global South results. In some cases, the findings may look less promising than those in the Global North because of differences in the timing of the studies and their publication status rather than contextual differences.

The evidence presented in this report thus reflects the most up-to-date experimental evidence on the effectiveness of a menu of misinformation interventions while also accounting for summaries of past findings. Table 3 displays each intervention organized by category, as well as the number of studies

collected based on the above criteria. We consider 155 unique studies that evaluate a total of 176 interventions (some studies are conducted in multiple countries and others test multiple interventions and thus span categories).

Table 3: Summary of intervention types

Category	Intervention	Studies Cited	Global North	Global South
Informational	Inoculation	25	18	7
Informational	Debunking	56	49	9
Informational	Credibility labels/tags	24	23	1
Informational	Contextual labels/tags	8	7	1
Educational	Media literacy	16	9	8
Socio-psychological	Accuracy prompts	13	11	3
Socio-psychological	Frictions	3	3	0
Socio-psychological	Social norms	14	11	3
Institutional	Platform alterations	10	8	2
Institutional	Politician messaging	4	4	0
Institutional	Journalist training	3	3	0

Once the studies were collected, team members tagged the country and year in which the study was conducted, whether it was preregistered or not, whether the study focused on marginalized populations or not, and whether the study identified heterogeneous treatment effects.

As shown above, the evidence base from the Global South is relatively thin compared to the Global North for most interventions. For some interventions, we lack any experimental evidence at all. We thus also conducted an expert survey that was designed to help us define a research agenda for the Global South in light of this disparity. The expert survey sample is composed of two groups: experts on misinformation and experts on governance interventions in the Global South.

Among each set of experts, we provided a list of the key interventions with a brief description of each. Survey participants were asked to allocate 100 dollars/points among a portfolio of interventions toward the ones they would expect to be the most successful in a Global South context. We also asked experts

to assess which interventions might be more or less effective in the context of the Global South than they had been found to be in the Global North. Both quantitative and qualitative data from the expert survey are analyzed in Section 5.

4. FINDINGS

4.1 INFORMATIONAL INTERVENTIONS

The first of the four intervention categories, informational interventions provide additional information about the factual basis of a claim, the credibility of its source, the origin of the claim, or the type of information presented (e.g., manipulated media). We identify four interventions in this category: inoculation, debunking, credibility labels, and contextual labels. These studies are often evaluated using survey experiments that are typically (though not always) conducted online.

4.1.1 INOCULATION

Inoculation, which is also known as prebunking, is a corrective intervention that occurs *prior* to an individual's exposure to a piece of misinformation. The application was first studied in the context of biological immunization and then, in the 1960s, analogized by McGuire (1964) as a way to protect against attempts at persuasion.

Inoculation is generally a two-step process. Individuals are warned against an imminent attack, introducing a sense of threat (forewarning), and then given a "dose" of the impending piece of misinformation to facilitate recognition along with counter-arguments to resist it (refutational preemption) (Compton et al., 2021). Inoculation interventions can focus either on correcting misinformation related to specific issues (e.g., climate change) or on raising awareness of the techniques commonly used to misinform people (e.g., providing testimony or evidence from fake experts). Inoculation interventions have been delivered through a variety of modes including written messages, videos, and online games. Regardless of their form, the underlying logic is that making individuals aware of their vulnerability to persuasion helps them to generate resistance. Our review of the evidence, detailed below, indicates that inoculation interventions are generally effective at reducing belief in misinformation (typically measured immediately after exposure) in both the Global North and Global South.

EVIDENCE ON INOCULATION FROM THE GLOBAL NORTH

Inoculation findings: Global North
Finding 1: Prebunking messages are generally effective in the short term.
Finding 2: Evidence is mixed on the effectiveness of technique inoculation.
Finding 3: Prebunking messages tend to produce moderate effects lasting at least one week.
Finding 4: Inoculation games, while innovative, may not be as effective as first thought.
Finding 5: In direct comparisons, inoculation is generally not as effective as debunking (discussed below).

Inoculation findings: Global North

Finding 6: In direct comparisons, inoculation is more effective than credibility labels and tags (discussed below).

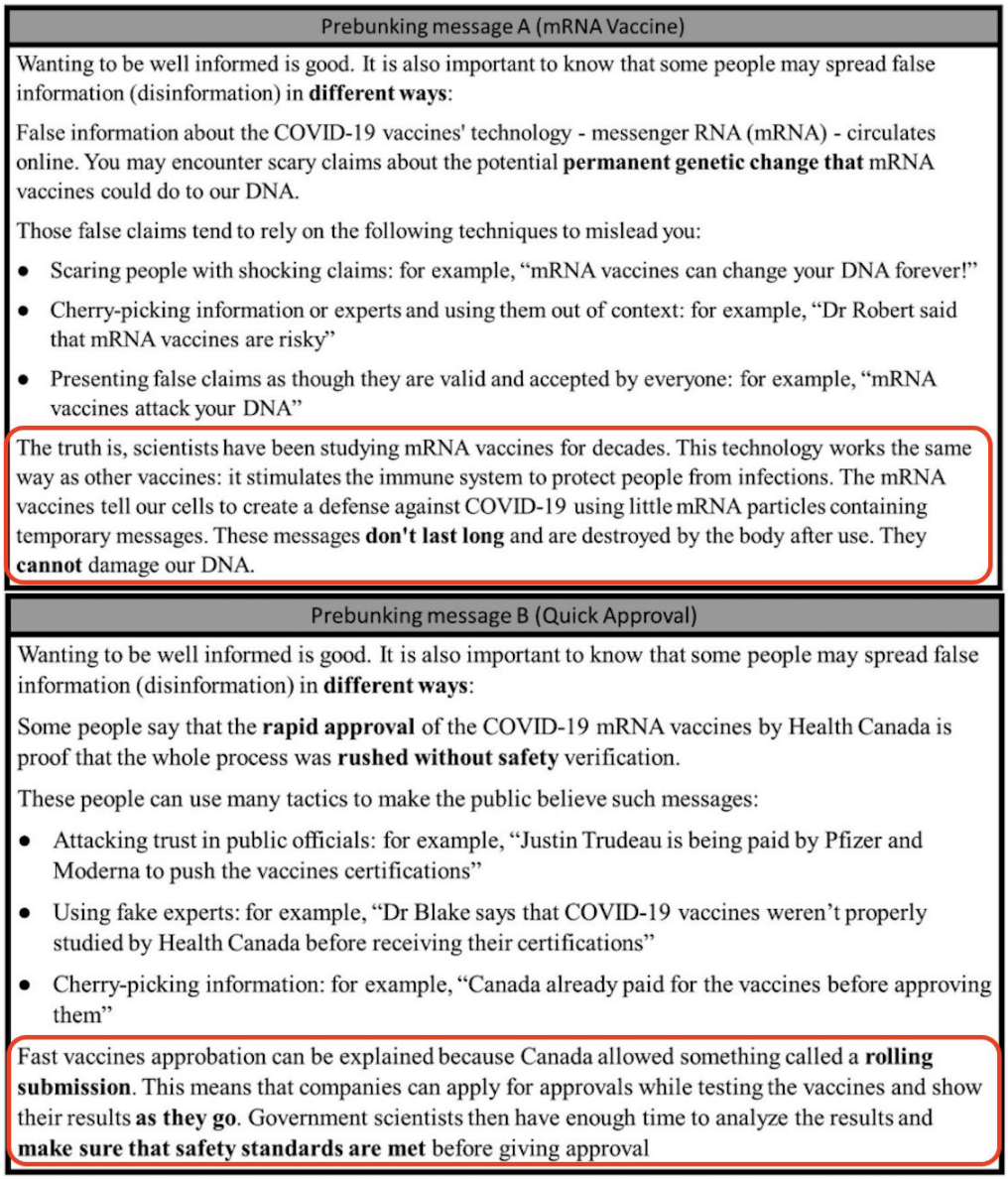
The majority of studies on inoculation find that it is effective for countering misinformation in the Global North. A recent review by Compton et al. (2021), which focused on evidence from Global North countries, reports that inoculation is effective at bolstering scientific confidence, reducing misinformed beliefs, and increasing self-reported behavioral intentions (e.g., greater likelihood to be vaccinated). Many early studies on inoculation report positive effects regarding climate change misinformation in particular. Several studies investigated the effects of pre- bunking misinformed claims in the well-known Oregon Petition, which famously stated there is no evidence that global warming is caused by human behavior and was signed by non-experts and fake online signatories (Kasprak, 2016). These studies find that both fact-based inoculation and technique-based inoculation were effective against this particular piece of misinformation. Although climate misinformation received a great deal of attention in their review, Compton et al. (2021) also report that inoculation was demonstrated to be effective at reducing misinformation about vaccines, biotechnology in agriculture, and animal research.

Specific examples of studies showing the effectiveness of inoculation in our evidence base include Vivion et al. (2022), which demonstrates that detailed pre- bunking interventions (see Figure 1) increased intentions to receive the COVID-19 vaccine among Canadians. However, the treatment had no effect on attitudes. Inoculation can also be effective at overcoming the persistent effects of misinformation even in the face of retraction or correction (Buczel et al., 2022), which is commonly referred to as the continued influence effect (CIE). One inoculation study conducted in Italy led to decreased perceptions of the plausibility of fake news among respondents with a high conspiracy mentality (Bertolotti and Catellani, 2023).

Not every study reports positive results, however. In one of the three null results studies included in this review, Jiang et al. (2022) find no effect of inoculation on COVID-19 vaccine attitudes or intentions among individuals in Hong Kong relative to a control group. Similarly, Schmid-Petri and Bürger (2022) find no evidence that inoculation preceding misinformation had any effect on climate change attitudes among German adults, showing that a previous study by Cook et al. (2017) did not replicate in the German context. Research also suggests that inoculation messages that have been commented on by others in an online environment did not influence smoking or COVID-19 attitudes (Dai et al., 2022). Nevertheless, the weight of the evidence suggests that inoculation works on average in the Global North.

With respect to technique inoculation, the evidence from the Global North is mixed. Technique-based inoculation interventions try to counter common approaches or strategies used to misinform people. These techniques include presenting information from fake experts, setting impossible expectations, and cherry- picking by only referencing evidence that supports selective claims (Cook et al., 2018). Four of the 19 inoculation studies reviewed here employ interventions aimed at countering such techniques. While some find that statements about false balance in scientific debates (Schmid et al., 2020) or more involved treatments such as inoculation games (Roozenbeek et al., 2022) are effective in reducing misinformed beliefs, other studies do not replicate the positive effect of informing individuals about fake experts on their beliefs about climate change (Schmid-Petri and Bürger, 2022). It may be the case that technique inoculation is not sufficient on its own to address specific misinformed claims.

Figure 1: Inoculation treatment from Vivion et al. (2022) (boxes added)



Prebunking message A (mRNA Vaccine)

Wanting to be well informed is good. It is also important to know that some people may spread false information (disinformation) in **different ways**:

False information about the COVID-19 vaccines' technology - messenger RNA (mRNA) - circulates online. You may encounter scary claims about the potential **permanent genetic change that** mRNA vaccines could do to our DNA.

Those false claims tend to rely on the following techniques to mislead you:

- Scaring people with shocking claims: for example, "mRNA vaccines can change your DNA forever!"
- Cherry-picking information or experts and using them out of context: for example, "Dr Robert said that mRNA vaccines are risky"
- Presenting false claims as though they are valid and accepted by everyone: for example, "mRNA vaccines attack your DNA"

The truth is, scientists have been studying mRNA vaccines for decades. This technology works the same way as other vaccines: it stimulates the immune system to protect people from infections. The mRNA vaccines tell our cells to create a defense against COVID-19 using little mRNA particles containing temporary messages. These messages **don't last long** and are destroyed by the body after use. They **cannot** damage our DNA.

Prebunking message B (Quick Approval)

Wanting to be well informed is good. It is also important to know that some people may spread false information (disinformation) in **different ways**:

Some people say that the **rapid approval** of the COVID-19 mRNA vaccines by Health Canada is proof that the whole process was **rushed without safety** verification.

These people can use many tactics to make the public believe such messages:

- Attacking trust in public officials: for example, "Justin Trudeau is being paid by Pfizer and Moderna to push the vaccines certifications"
- Using fake experts: for example, "Dr Blake says that COVID-19 vaccines weren't properly studied by Health Canada before receiving their certifications"
- Cherry-picking information: for example, "Canada already paid for the vaccines before approving them"

Fast vaccines approbation can be explained because Canada allowed something called a **rolling submission**. This means that companies can apply for approvals while testing the vaccines and show their results **as they go**. Government scientists then have enough time to analyze the results and **make sure that safety standards are met** before giving approval

A central question for any information intervention is whether effects endure beyond immediate exposure to the corrective information. Of note, inoculation messages tend to produce moderate effects that last at least one week. Compton et al. (2021) note in their review that inoculation effects do not decay significantly after one week when used to address climate change and health misinformation. Of those studies investigating standard inoculation interventions in our database (i.e., not game-based inoculation), only one examines the outcomes of interest at least one-week post-intervention exposure in the Global North. Brashier et al. (2021) find that prebunking increases discernment by 6–7% one-week post-exposure (a weaker effect than debunking, which increased discernment by approximately 25%), providing evidence consistent with prior research that prebunking effects decay relatively little over time.

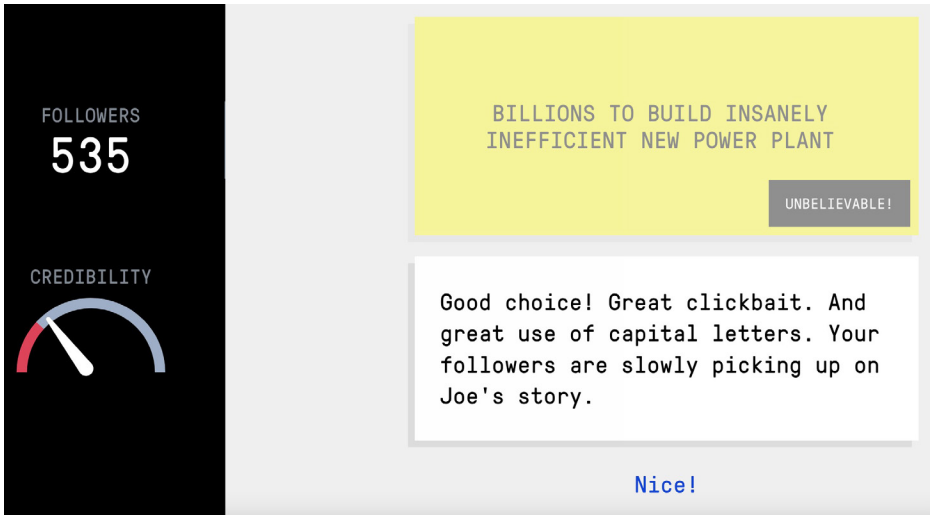
While there is evidence that both issue-based and technique-based inoculation work, inoculation *games*, while innovative, may not be as effective as originally thought. Indeed, there is a sizeable literature dedicated to evaluating the effective- ness of inoculation games, a specific type of intervention that combines elements of an online game, technique inoculation, and media literacy (see Figure 2). The two most widely known inoculation games, *Bad News* and *Go Viral!*, put the individual in the role of a misinformation distributor to reveal the techniques and tricks that

would be employed in a real-world setting. The game designers hope to demystify the misinformation process by providing a behind-the-scenes setting for individuals to explore how easy it is to mislead, lie, and provide inflammatory content for the sake of engagement.

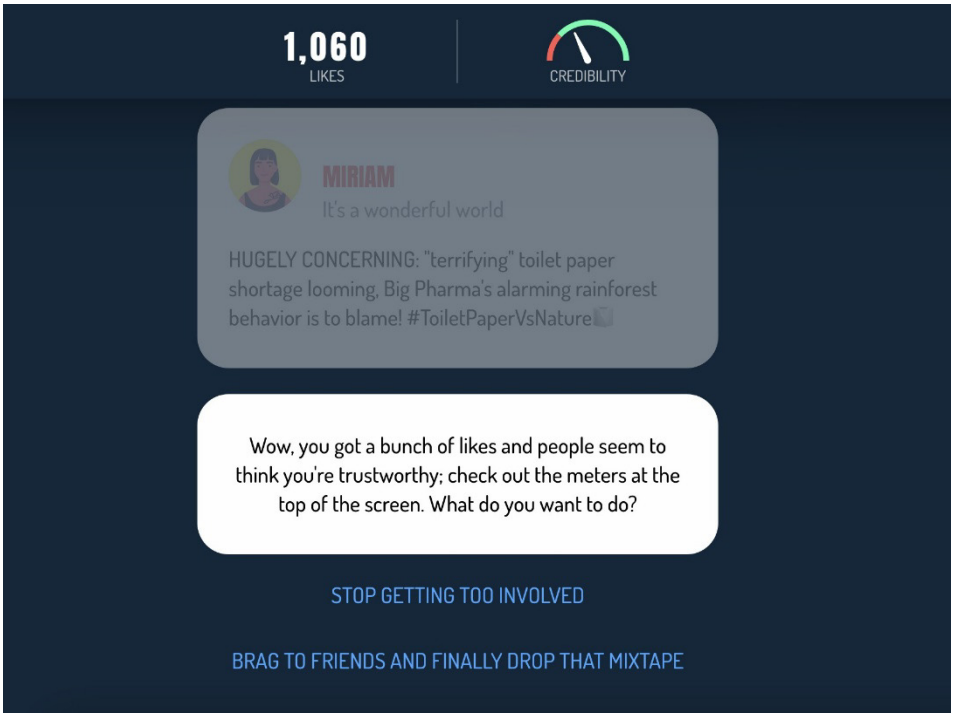
Until recently, the evidence suggested that games such as these were successful in combating misinformation at the belief stage (Roozenbeek et al., 2021; Maertens et al., 2021; Basol et al., 2021). However, Modirrousta-Galian and Higham (2023) reanalyze five of the most cited inoculation game studies and find four out of five actually show no effect on helping individuals discern true news from misinformation, with the fifth study producing inconclusive results. Their apparent effectiveness at reducing belief in misinformation is the result of them causing people to rate *all* news as false more frequently.

A final important consideration is how well prebunking works relative to (or in conjunction with) other informational interventions. One intervention of particular interest is debunking, which provides corrective information *after* exposure (discussed below). Some have hypothesized that prebunking might be more effective than debunking because the effects of exposure to misinformation can be hard to undo afterward (Cook, 2016). While these two interventions have generally been studied in isolation, a small number of recent studies have examined them together. The existing evidence suggests that inoculation is generally not as effective as debunking at addressing misinformed beliefs and behavioral intentions. Vraga et al. (2020) offer one of the first attempts to examine prebunking and de- bunking in tandem, looking at the effect of providing fact- versus logic-focused (technique-focused) messages both before and after exposure to misinformation about climate change. They find that technique-based strategies work for both prebunking and debunking, although fact-based strategies work only for debunking. While debunking was more effective at reducing misperceptions than prebunking, it was equally effective at reducing credibility of misinformation. Brashier et al. (2021) also find that, when comparing inoculation to both debunking and credibility labels, debunking was most effective at improving individuals' ability to discern between true and false news. Finally, Li et al. (2022) find that debunking was more effective than prebunking at reducing intent to promote misinformation, although both reduced individuals' reliance on misinformation.

Figure 2: Screenshots from inoculation games (Maertens et al., 2021; Basol et al., 2021)



a) Bad News



b) Go Viral!

One study uses prebunking and debunking in tandem. In a study of unvaccinated adults in the U.S., Amazeen et al. (2022) find the effect of inoculation varies depending on whether individuals are also exposed to debunking messages and have skeptical attitudes about COVID-19 vaccines. While general and specific inoculation messages had positive effects on those who already supported the COVID-19 vaccine, they made people more skeptical about the COVID-19 vaccine when combined with debunking and administered to individuals who were already skeptical.

While inoculation does not seem to be as effective as debunking, it may be more effective than other informational interventions discussed in this report. A handful of studies have examined the relative effectiveness of prebunking and credibility tags (discussed below). Two studies find that prebunking is more effective than credibility tags (McPhedran et al., 2022, 2023), which provides support for the effectiveness of inoculation versus other strategies apart from debunking.

EVIDENCE ON INOCULATION FROM THE GLOBAL SOUTH

Inoculation findings: Global South

Finding 1: Inoculation interventions tend to be effective on average, especially when implemented over long periods of time in partnership with local news entities.

Finding 2: As in the Global North, gamified inoculation interventions do not durably increase discernment.

As in the Global North, inoculation also tends to work in the Global South, especially when interventions are implemented over long periods of time and in partnership with local news entities. This report includes seven studies conducted in the Global South; four were conducted in India and the others took place in Brazil, South Africa, and China. While there are far fewer studies conducted in the Global South than in the Global North, four of the seven are large-scale field experiments conducted over an extended period of time. Three of those four find that inoculation generally works. The remaining three are game-based interventions and provide more mixed results. Nevertheless, the evidence suggests that, on balance, inoculation interventions are effective in the Global South.

Three large-scale field experiments provide evidence that inoculation works in the Global South (Bowles et al., 2023; Pereira et al., 2022b; Garg et al., 2022). Bowles et al. (2023) employ a six-month long treatment in which individuals received biweekly fact-checks via WhatsApp through a partnership with a South African fact-checking organization. The fact-checks covered a variety of topics, including COVID-19, other health-related topics, politics, and high-salience cultural content. The specific fact-checks also varied in length, written versus audio delivery, and entertainment versus empathy messaging. All treatment types succeeded in increasing discernment between true and false news, but the study finds that simple textual messages and an empathetic podcast were particularly effective. In sum: "...repeated, short, and sharply-presented factual proclamations from a credible source are more likely to train people to approach information more critically than longer-form edutainment, unless such content goes out of its way to empathize with consumers. The combined implication is that short but empathetic fact-checking may be the most effective means of inoculating people against misinformation" (23).

Pereira et al. (2022b) conduct a field experiment in São Paulo to investigate the effectiveness of providing six-month newspaper subscription vouchers and fact-checking emails to participants, working. They find that their intervention, which was carried out in collaboration with Brazil's main newspaper, reduced belief in false news without increasing skepticism towards true news, possibly due to the treatment increasing access to resources about false news, internal motivation towards truthfulness, and/or knowledge about how to identify false news generally. Similarly, Garg et al. (2022) shows that inoculation works in India, although their findings differ in one important way from Pereira et al. (2022b). Garg et al. (2022) provided weekly fact-checks and narratives regarding salient and politically relevant targets of misinformation on a bespoke phone application. The intervention was successful at increasing discernment,

although it also caused minor increases in skepticism about true news unlike in Pereira et al. (2022b). Iyengar et al. (2022) also provide evidence of the effectiveness of inoculation in India using an online game-based intervention, although this study was conducted with a student sample and the effects were small.

In contrast to the above, a large-scale field experiment in India by Badrinathan (2021) tests a single, hour-long media literacy training (see more below) with an inoculation component. However, after two weeks, no effects were found (outcomes were not measured directly after treatment). The author speculates that a single training may not have been sufficient to address the enormous influence of misinformation, especially in states with less Internet penetration and lower digital literacy. Since the treatment was complex and included elements other than inoculation, it is difficult to draw conclusions about the specific effects of inoculation from this study.

While Iyengar et al. (2022) finds support for inoculation using an online game intervention, additional studies conducted in China and India find no effect. Three of the seven studies conducted in the Global South employ an online game intervention. One study investigates the effect of gamified inoculation in north India (Harjani et al., 2023) with the bespoke game *Join this Group*, which places individuals in the position of an undercover investigator into misinformation (as opposed to an up-and-coming distributor of it, the role typically employed in online inoculation game interventions). The authors find no effect and speculate that a combination of cultural factors (such as digital literacy and rural context) and experimental design factors contributed to the null result. Iyengar et al. (2022) test the effectiveness of the *Bad News* game discussed above among a sample of Indian students. Although the authors find that exposing individuals to impersonation and conspiracy theory techniques increases discernment, the effects are small.

Likewise, a recent study in China demonstrated that an online inoculation game implemented via WeChat was initially effective, but the effects did not persist after one week (Ma et al., 2023). As in the Global North, there is little convincing evidence that gamified inoculation is effective in the Global South.

In summary, inoculation has been shown to be an effective method to address misinformation. Those interested in implementing an inoculation intervention may wish to employ the intervention repeatedly, as more exposure to correct claims may increase effectiveness. Logistically, inoculation may be difficult to implement given that it relies on knowing about future exposure to a specific false or unsupported claim so people can be warned in advance. By contrast, the evidence suggests that inoculation games, while already developed and easily employable at scale, cause individuals to doubt *all* news (not just false news) and thus are not an effective alternative to traditional inoculation.

4.1.2 DEBUNKING

Debunking is the correction of a specific false or misleading claim after exposure. The aim of debunking is to undo or reverse the effects of a particular piece of misinformation. Debunking can take many forms, including fact-checking, algorithmic correction on a platform, and/or social correction by an individual or group of online peers. It can also vary in its level of specificity, mode of delivery, and timing.

Debunking is one of the most widely studied misinformation interventions. As such, several debunking-specific reviews have been published recently. In one meta-analysis, Chan et al. (2017) review 20 separate experimental studies from 1994 to 2015 and find that debunking is widely effective, although

they note that detailed debunking messages are more effective than general fact-checks. A recent literature review (Lee and Shin, 2021) provides updated support for the efficacy of debunking while also noting the importance of context-specific moderators, such as an individual's prior attitudes and the timing of the correction relative to an individual's encounter with misinformation. According to Lee and Shin (2021),

immediate corrections are more effective than delayed ones, and corrections in general are more effective when they support an individual's pre-existing beliefs. In a more recent meta-analysis, Walter et al. (2020) examine 30 unique experiments from 20 studies on fact-checking. Although they confirm the overall positive effects claimed by Chan et al. (2017) and Lee and Shin (2021), they also note that effect sizes weaken as the intervention more closely resembles a real misinformation encounter. Additionally, they discuss differing levels of effectiveness depending on one's ideological or partisan leaning (i.e., Republicans/conservatives may be less receptive to fact-checks than Democrats/liberals).

EVIDENCE ON DEBUNKING FROM THE GLOBAL NORTH

Debunking findings: Global North

Finding 1: Debunking is generally effective at correcting misinformed beliefs.

Finding 2: Debunking generally seems to be more effective at correcting misinformed beliefs (e.g., that vaccines cause autism) than at inducing behavioral change (e.g., the decision to get vaccinated).

Finding 3: Corrections from experts and official sources (e.g., the CDC in the U.S.) seem to be more effective than corrections by other people.

Finding 4: Findings are mixed as to whether objective corrections (i.e., corrections based on facts) are more, less, or equally effective as subjective ones (i.e., corrections based on emotion or personal experience).

Finding 5: There is little evidence that corrections backfire, even among subgroups who are especially prone to believe a specific piece of misinformation.

Finding 6: There is little evidence that the specific format of the correction (e.g., whether or not it includes an image, whether or not it involves humor, whether or not it repeats the specific piece of misinformation that is being debunked, etc.) matters for its efficacy.

Detailed debunking messages are effective at reducing misinformed beliefs and (to a lesser extent) altering behavioral intentions. Most recent debunking research in the Global North has focused on misinformation related to health, especially COVID. This recent focus on health is important, as Vraga et al. (2019) find that debunking is more effective for health-related misinformation than for other forms of misinformation (e.g., misinformation about climate change or gun control). But the efficacy of debunking is not strictly limited to health; indeed, corrections have been shown to be effective at reducing beliefs in various types of falsehoods by politicians (Aird et al., 2018), neuroscience myths (Smith and Seitz, 2019), and false claims about immigration (Hameleers and van der Meer, 2020).

Debunking generally seems to be more effective at correcting misinformation than at inducing change in behavioral intentions (Porter et al., 2023; Vraga et al., 2021), though debunking interventions have been found to increase vaccine intentions (Schmid and Betsch, 2019) and self-reported compliance with preventative health measures (van der Meer and Jin, 2020) in some cases. Unfortunately, with only a handful of exceptions (e.g., Clayton et al., 2021; Dai et al., 2021), few studies measure whether behavioral intentions translate into actual behavior (e.g., whether individuals who express an intention to get vaccinated actually do so).

Most recent debunking research has found that corrections from experts or official sources are more effective than corrections from other people. Some studies find that expert corrections (e.g., from the CDC) are more effective than social ones (van der Meer and Jin, 2020; Vraga and Bode, 2017). Expert corrections may be especially effective if the experts' judgments have high "social endorsements" (e.g., if they receive many "likes" or shares on social media) (Wang, 2021). Other studies find that algorithmic and social corrections are equally effective when they are substantiated by media reports and expert judgments (Bode and Vraga, 2018). A subset of debunking interventions involve the use of "technique rebuttal." Technique rebuttal is distinct from technique inoculation (discussed above), and involves exposure to a false or misleading claim followed by a correction of the logic or rhetorical strategy used to make the claim.[8] Technique rebuttal has been found to be effective at reducing misinformed beliefs, especially among populations with low levels of scientific confidence, but more cross-national evidence is needed. Schmid and Betsch (2019) find that both rhetorical rebuttals and claim corrections are effective at countering misinformation about vaccines and climate change in the U.S. and Germany.

Debunking has withstood several critiques. For instance, some have speculated that the repetition of misinformation in corrective messages does more harm than good due to the continued influence effect (Thorson, 2016) — a theory positing that misinformation remains in one's memory and influences one's thinking even after it has been debunked. However, empirical tests of this claim (Ecker et al., 2017, 2020) find not only that corrections work, but also that they tend to work best when they include an explicit reminder about the misinformation even among individuals who were not previously exposed to the false or misleading claim (Ecker et al., 2020). Others have suggested that debunking is subject to backfire effects, whereby individuals react to corrections by doubling down on their belief in the falsehood being corrected (Nyhan and Reifler, 2010). However, numerous recent studies have found no evidence of backfire effects (Wood and Porter, 2019; Schmid and Betsch, 2019; Nyhan et al., 2020; Porter and Wood, 2021), lending credibility to the continued use of corrective messages.

Finally, there is little evidence that the specific format of the corrections matters for its efficacy. For example, while narrative corrections (i.e., corrections that incorporate a story or emotional element) can be effective in altering attitudes (Sangalang et al., 2019), they appear to be no more or less effective than non-narrative corrections at reducing false beliefs (Ecker et al., 2020). Similarly, multi-modal fact-checks (combining text and visuals) appear to be no more less effective than purely textual corrections, even in response to misinformation that is presented in multi-modal fashion (Hameleers et al., 2020).

[8] For instance, skeptics of vaccines sometimes argue that vaccines should be proven to be 100% safe before they are widely administered. Debunking this claim would involve providing detailed evidence about the generally excellent safety record of vaccines. Technique rebuttal would focus on the fact that no medical product or procedure is ever 100% safe, and that to expect otherwise is unreasonable.

EVIDENCE ON DEBUNKING FROM THE GLOBAL SOUTH

Debunking findings: Global South

Finding 1: As in the Global North, debunking is generally effective at correcting misinformed beliefs, though there is some variation across contexts and types of misinformation.

Finding 2: Corrections from sources that share personal, political, or religious ties with the recipient generally appear to be more effective.

Finding 3: As in the Global North, evidence is mixed regarding the effectiveness of debunking on behavioral intentions.

The vast majority of debunking research has focused on the Global North. However, a growing group of debunking studies focus largely or exclusively on countries in the Global South. This report considers nine individual studies that examine debunking interventions in 10 Global South countries: Argentina, Brazil, India, Indonesia, Nigeria, Pakistan, Peru, Sierra Leone, South Africa, and Zimbabwe. All are quite recent (the oldest was published in 2020).

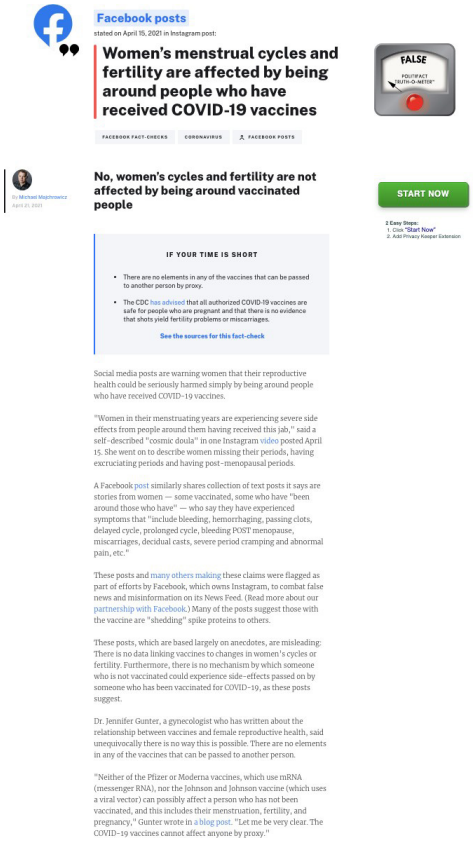
As in the Global North, debunking has been shown to be effective at reducing beliefs in false or misleading claims in the Global South, though there is some variation across contexts and types of misinformation. Some scholars have found debunking to be just as effective in Global South contexts as in the Global North. Porter and Wood (2021), for example, administered fact-checking interventions in four countries, three of which are in the Global South. The fact-checks covered a range of topics, including the economy, crime, COVID-19, and local politics. In total, 22 fact-checks were provided, five for each country and two that were presented to the entire sample. The authors find that fact-checking reduced beliefs in misinformation in Argentina, Nigeria, South Africa, and the U.K. Additionally, the effectiveness of the fact-checks endured at least two weeks after exposure, providing promising evidence for the potential durability of debunking interventions. However, other scholars have noted the varying effectiveness of debunking across disparate Global South settings or distinct categories of misinformation. Carey et al. (2020) report mixed results from corrective public health messages about the Zika virus in Brazil: "overall, across two Zika experiments, the myths correction treatment measurably decreased belief in 7 of 13 statements about Zika, including 6 of the 9 accurate statements that were tested" (7). That is, not only did debunking fail to reduce beliefs in all false or misleading claims about Zika, but it reduced beliefs in some true statements as well. The authors note that debunking was more effective at countering misinformation about yellow fever, possibly because yellow fever has been present in Brazil for much longer than Zika. They speculate that misinformation around novel threats may be more difficult to correct.

Porter et al. (2023) report similarly mixed results. They use fact-checking to correct misinformation about COVID in 10 countries (see Figure 3 for examples), six of which are in the Global South. Participants were exposed to three pieces of misinformation, two global (e.g., that the vaccine alters DNA) and one country-specific (e.g., that Nigeria bought low-quality vaccines). Fact-checking was quite effective overall but its efficacy varied across countries. Only in the U.S. did fact-checking successfully reduce beliefs in *all* false or misleading claims; in contrast, fact-checking had no statistically detectable effect on misinformed beliefs in Indonesia. Corrections of country-specific misinformation were

ineffective in Brazil, Nigeria, and South Africa; in Peru and India, by contrast, corrections of global misinformation were ineffective. The authors consider several possibilities that might account for this discrepancy across contexts, including differences in the capacity of local fact-checking entities, the content of misinformation across countries, how outcomes were measured, and how difficult it was to correct different pieces of misinformation.

Pereira et al. (2022a) are the only debunking study conducted in the Global South in our data that reports null results. They find that fact-checking false rumors regarding the 2018 Brazilian election did not reduce belief in both political and nonpolitical rumors. The null findings were not explained by political interest, prior beliefs, or socio-demographic characteristics. The authors speculate that social endorsements may strengthen third-party fact checking, especially in contexts with low media and digital literacy.

Figure 3: Debunking treatments from Porter et al. (2023)



(a) U.S. stimulus

(b) Brazil stimulus

Recent research often focuses on debunking misinformation from WhatsApp, which has become an important vector of misinformation in Global South countries. These studies have also experimented with a variety of audio and video formats for delivering corrections, which may be more effective for reaching marginalized populations where literacy rates tend to be low. Bowles et al. (2020), for example, partner with two non-governmental organizations in Zimbabwe to show that debunking increased recipients' belief in correct claims about COVID-19 and strengthened their intent to adhere to a nationwide lockdown. Other studies take a similar approach. In a study focused on Sierra Leone, Winters et al. (2021) find that detailed (rather than general) corrective audio dramas reduced beliefs in misinformation about typhoid and malaria circulated on WhatsApp. As in the Global North, this intervention was more effective when it described the specific piece misinformation before debunking it. Finally, Badrinathan and Chauchard (2023) find that social corrections in WhatsApp group chat conversations success- fully reduced Indian participants' belief in six of seven pieces of misinformation. They also find that corrections were no more effective when they explicitly listed a source and that simple corrections were just as effective as detailed ones (contrary to Winters et al. 2021).

Recent research also suggests that debunking may be more effective in Global South contexts when corrections are delivered or endorsed by an individual or group that shares an identity with or social tie to the consumer. For example, in one study focused on Indian slum residents, Armand et al. (2021) combine debunking and social norm-related interventions (discussed later in this report) in audio and video messages on WhatsApp. The corrections were delivered by a doctor, who in turn was introduced by either a Hindu or a Muslim speaker.[9] The corrections successfully reduced belief in the false claim that vegetarian diets prevent COVID, but only when the speaker and recipient shared the same religious identity. The corrections had no effect on the mistaken belief that Indian immune systems are uniquely resilient to COVID, but they did increase behavioral intentions to mask, social distance, and follow safety guidelines.

Consistent with Armand et al., Pasquetto et al. (2022) find in India and Pakistan that corrections were more likely to be shared by recipients who had social or political ties in common with the individuals sending them. The authors also find that corrections received as audio files were more effective than text- or image-based messages.

Importantly, there are relatively few debunking studies in the Global South that investigate topics besides health — an important limitation of the existing literature and avenue for future research.

Overall, both past and recent literature on debunking demonstrate its efficacy at combating misinformation. The most effective debunking interventions tend to provide more details, cite expert sources, and/or contain endorsements from in-group members. Prior research suggests that the format of debunking messages (i.e., as narrative, text, or visual) does not seem to change their effectiveness systematically. However, significant implementation challenges remain. Like inoculation, debunking addresses specific misinformed claims or arguments and is thus difficult to scale given the enormous amount of misinformation that is produced daily around the globe. Moreover, much of the work on debunking has been performed in stylized settings such as online surveys and simulated social media environments in which people are exposed directly to the messages, suggesting uncertainty about how effective debunking is in the real world where reaching audiences is more difficult. Finally, some false

[9]Religious identity was cued by style of dress, greeting phrase, and speaker name.

beliefs may be more deeply ingrained than others, so implementers should think carefully about how responsive people are likely to be to corrective information in their context.

4.1.3 CREDIBILITY LABELS AND TAGS

Rather than providing detailed corrections, credibility labels and tags provide cues about the truth value of a piece of content without offering explicit explanations. These labels and tags are often attached to the presentation of a misinformed claim, which differs from prebunking and debunking messages presented before and after exposure, respectively. These interventions are relatively recent and became more common online after the 2016 U.S. presidential election. For in- stance, Facebook implemented their plan for "disputed" tags in 2016 and Instagram began including fact-checked tags in 2019 (Guynn, 2016; Meta, 2019). Like- wise, Twitter announced in early 2020 that they had plans to add their own tags of "misleading," "disputed," or "unverified" to flagged tweets (Roth and Pickles, 2020).

EVIDENCE ON CREDIBILITY LABELS FROM THE GLOBAL NORTH

Credibility labels findings: Global North

Finding 1: Credibility labels work on average.

Finding 2: Labels are most effective when they provide justification for their presence, provide clear true or false ratings, and reference expert fact-checkers.

Finding 3: The duration of credibility label effects may be short, but the evidence is mixed.

Finding 4: Credibility tags and labels may have unintended consequences for untagged content.

A number of studies suggest that credibility labels and tags work in countering misinformation. Researchers have examined a wide variety of tag and label types, including those that simply label information as "disputed" as well as those that label information more definitively as "true" or "false". Overall, the literature suggests that "disputed" tags reduce the credibility of targeted claims, especially when tags are accompanied by additional information on the source or explanation (see Figure 4 for one example). Specific studies show that tagging information as "disputed" and including sources were effective at decreasing perceived credibility of fake news on Facebook (Mena, 2020), increasing discernment between true and false news (Seo et al., 2019), and reducing intentions to share misinformation (Mena, 2020; Yaqub et al., 2020; Celadin et al., 2023). Similarly, Kirchner and Reuter (2020) find that "disputed" tags with explanations about why posts were disputed were more effective at increasing discernment than simple "disputed" tags, which suggests that pairing tags with debunking may be more effective than independently implementing either intervention.

Figure 4: Disputed tags from Mena (2020)

Other studies find even greater support for the effectiveness of tags that explicitly label information as true or false. For instance, Amazeen et al. (2018) show that using visual truth scales to fact-check might increase discernment between true and false statements, while Clayton et al. (2020) find that definitive tags stating a claim was "rated false" are more effective than "disputed" tags.

The evidence also suggests that credibility labels by both professional fact- checkers and ordinary people are effective, although those by professional fact- checkers are especially impactful. Kim et al. (2019) test the effects of expert fact-checks, user ratings of the content, and user ratings of the information source on beliefs about misinformation in an online U.S. sample. They find that presenting expert and other users' ratings of headlines led to decreased believability, but only when the ratings given to the headlines were low (as opposed to medium or high). There was no effect when other users rated the source of the headline as opposed to the content, and these effects did not extend to reading, liking, commenting, or sharing intentions. Their findings on the successful effects of expert fact-checkers are consistent with Celadin et al. (2023), who demonstrate that ratings of source trustworthiness affected the propensity of users to share misinformation, especially strong when the labels were from professional fact-checkers. Seo et al. (2019) responds to growing interest in using computational methods to detect false news by testing the effects of fact-check labels from humans versus a machine-learning algorithm. Interestingly,

while computational methods did a better job of detecting misinformation, users trusted human fact-checking more – a finding with important implications for whether computational methods can ever become a desirable alternative to human fact-checking.

We note two important caveats to these positive findings, however. First, the effects of credibility tags may vary for different kinds of individuals. For instance, in the context of the U.S., effects may vary by party identification — definitive tags may be more effective for Democrats than Republican or Independents, especially if the fact-checker is not perceived to be ideologically congruent (Jennings and Stroud, 2021). However, Amazeen et al. (2018) note that their visual rating scales were effective even when the correction was politically uncongenial.

Second, the effects of fact-checking could be short-lived, though the evidence is mixed. Only two of the 24 Global North studies examine outcomes beyond the immediate study environment. A study by Grady et al. (2021) that tested three versions of "fake-news" labels on a U.S. sample found large effects on beliefs in the short-run but they had largely faded for the exact same articles two weeks later. In contrast, Brashier et al. (2021) find that labeling information as "true" or "false" successfully affected accuracy ratings one week after exposure.

It is important to note that several studies reviewed here found mixed or no effects of credibility labels. One found that the use of "disputed" labels had no effect on perceived agreement with fake headlines (Gao et al., 2018), while another that noted that "disputed" tags on Twitter only reduced sharing intentions among Democrats and Independents (Lees et al., 2022). A series of studies by McPhedran et al. (2022) and McPhedran et al. (2023) found that labeling misinformed Facebook posts as false was not effective at reducing "likes/loves" compared to inoculation. Finally, researchers have noted the potential downsides of credibility labels.

Utilizing credibility labels could result in an increase in the perceived veracity of *untagged* information that could still be false or misleading (Seo et al., 2019; Pennycook et al., 2020). That is, the use of "disputed" tags or ratings in some instances could have the unintended effect of making false untagged headlines appear to be accurate. Another potential downside has to do with what happens when true information is erroneously tagged as false. As Freeze et al. (2021) show, wrongful tagging can lead individuals to ignore truthful information.

EVIDENCE ON CREDIBILITY LABELS FROM THE GLOBAL SOUTH

Credibility label findings: Global South
Finding 1: One study finds that credibility labels decrease individual sharing of misinformation; however, more studies are needed.

Our search identified only a single randomized controlled trial that has tested the effectiveness of credibility labels in the Global South. The study, by Nekmat (2020), examines the effectiveness of credibility labels in Singapore. The credibility labels take the form of fact-check alerts that provide a simple warning when information has been disputed by third-party fact-checkers or news publishers. Exposure to these fact-checking nudges resulted in lowered sharing intentions, especially when attached to a mainstream news source as opposed to a non-mainstream source. While this study suggests that credibility labels can also work in the Global South, more studies are needed to form firm conclusions.

Ultimately, the literature on credibility labels provides three considerations for increasing their efficacy. First, labels should clearly rate information; "disputed" tags or ratings are not as effective as "true" or "false" labels. Next, those intending to use credibility labels should also include a brief but explicit justification for why information was marked true or false. Lastly, those labels and tags which include a source were seen as more credible than those lacking that additional context.

4.1.4 CONTEXTUAL LABELS AND TAGS/PROVENANCE CUES

As audiovisual misinformation becomes more common, contextual labels and tags and provenance cues have become an increasingly important type of intervention. Contextual labels and tags provide additional information to help consumers understand and contextualize a particular piece of (mis)information. Like credibility tags, they are a fairly recent phenomenon. Twitter began labeling manipulated media in 2020 (Roth and Achuthan, 2020). More recently, Twitter rolled out their Community Notes function, which allows users to offer additional information or context about a claim without necessarily claiming that a tweet or the information therein is false (Twitter, 2023).

Contextual labels can be applied to written, verbal, or visual misinformation. By contrast, provenance cues provide media-specific information about the source or alteration of a picture, video, and/or audio clip. Like contextual labels, provenance cues provide additional details without directly addressing the veracity of the underlying content. These innovations appear to be popular, but consumers tend to prefer that they be accompanied by explanations for their use. Sherman et al. (2021), for example, employ a mixed-methods approach to demonstrate that, when consulted, users expressed a desire for clear statements, such as "confirmed" or "un- confirmed," as well as justifications for why particular images received these labels.

EVIDENCE ON CONTEXTUAL LABELS FROM THE GLOBAL NORTH

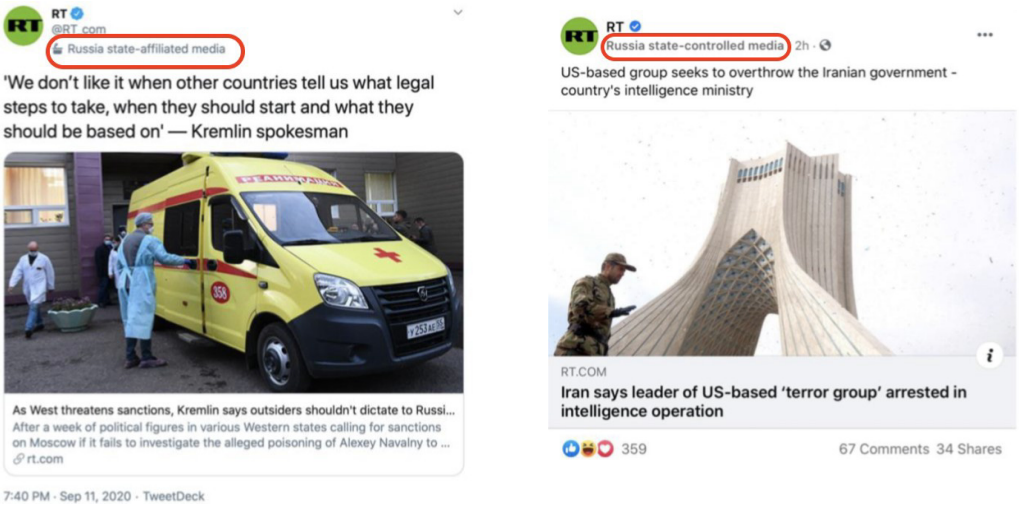
Contextual labels findings: Global North
Finding 1: The use of contextual labels and provenance cues may reduce belief and sharing intentions, but the evidence is mixed, and more is needed.
Finding 2: Verified badges attesting to the authenticity of a source do not affect perceptions of credibility or sharing intentions.

The use of contextual labels and provenance cues in the Global North may reduce misinformation beliefs and sharing intentions, but results are mixed and more evidence is needed. Only a handful of studies have investigated the impact of contextual labels and provenance cues in Global North settings. Nassetta and Gross (2020) demonstrate that labeling messages as originating from state-controlled media (see Figure 5 for examples) increased concern about fake news and its potential effects on election outcomes. The labels also helped counteract the negative effects of election misinformation, such as reduced trust in mainstream media. Bereskin (2023) tests the effects of provenance cues and traditional fact- checking. She provides suggestive evidence from a pilot study that fact-checking is more effective at countering climate change misinformation, while still noting the potential for provenance cues when paired with other interventions.

Contextual labels have also been used to attempt to mitigate the potential harms caused by "deepfakes" — computer-generated videos of people saying or doing things they have not said or done in reality. Individuals struggle to distinguish between real videos and fake ones, and at least one study finds that warning about the existence of deepfakes decreased trust in the veracity of *all* sorts of (video) content whether real or fabricated (Ternovski et al., 2021).

Relatedly, verified badges that attest to source authenticity do not seem to have an effect on perceptions of credibility or sharing intentions. This distinction is important because verified badges are generally attached to accounts of public interest and provide information about whether or not the source is *authentic* and not whether it is *credible*. Prior to Elon Musk acquiring Twitter in late 2022, a handful of studies reported that the blue verification checks failed to increase perceptions of news or tweet credibility (Edgerly and Vraga, 2019; Vaidya et al., 2019).

Figure 5: Context labels from Nassetta and Gross (2020)



a) Twitter state media label (b) Facebook state media label

Finally, Chakroff and Cole (2023) recommend using reverse-image searching to improve image perceptions. A type of provenance cue, reverse-image searching allows individuals to find the earliest known version of an image and retrieve information about the publisher or photographer. The proposed tool tested in the study would present individuals with original, unedited photos that they could then compare to images disseminated on social (or other) media, allowing them to judge whether the images are deepfakes or not. However, preliminary results are mixed. Individuals armed with reverse-image searching were less likely to share images that they categorized as "Iffy," but no more or less likely to share either "Fake" or "Real" images. In qualitative analysis, the researchers found that individuals were still willing to share fake images in part because they found them amusing or wanted to display a cool edit, not necessarily because they believed them to be real. Notably, those with higher digital literacy were more affected by provenance cues of this nature.

EVIDENCE ON CONTEXTUAL LABELS FROM THE GLOBAL SOUTH

Contextual labels findings: Global South

Finding 1: One study suggests that these labels could lower perceived credibility of misinformation in the Global South. However, the evidence base is extremely limited.

To our knowledge, only one randomized controlled trial has evaluated contextual labels in a Global South setting. Tandoc et al. (2022) report on a very simple intervention implemented on WhatsApp: the forwarded tag. The authors note that a forwarded tag is quite salient on WhatsApp due to the lack of other contextual information attached to forwarded messages. The authors find that the presence of a forwarded tag reduced the credibility of the tagged message among WhatsApp users in Singapore.[10] But this is just one study focused on just one platform-specific intervention. As a result, we cannot reliably judge the efficacy of contextual labels or provenance cues in Global South countries.

Overall, evidence on contextual labels is thin. However, some factors have been identified that may affect the degree of efficacy in combating misinformation. While verified badges do not seem to affect perceptions of news credibility, providing source information may cue individuals about whether the information they are consuming is trustworthy or deserving of additional scrutiny.

4.2 EDUCATIONAL INTERVENTIONS

Although seemingly similar to informational interventions, educational interventions provide broad information designed to build one's skills in consumption of media and are not typically related to specific information content. We consider only one intervention in this category: media literacy, which also includes digital literacy interventions. Studies of media literacy are often tested using online sur- vey experiments but field experiments have also been conducted online, via text messaging, and in in-person settings.

4.2.1 MEDIA LITERACY

Media literacy interventions provide participants with tools or tips for identifying common types of misinformation tactics. Ultimately, the goal of these interventions is for individuals to develop broad and long-lasting skills and competencies to discern true from false claims in future encounters with new information. The primary assumption is that individuals do not have the necessary educational and skill level to navigate a complex traditional and social media ecosystem where misinformation and distorted news abound.

Media literacy interventions therefore seek to provide individuals with skills and approaches that will help them identify dubious claims and misinformation tactics. These trainings come in many forms, including workshops, short videos, one-on-one lessons, and infographics. The extensiveness of the training also varies considerably from a minute or two to eight weeks (see Table 4). Thus, it is challenging to compare across studies, as each typically measures the impact of only one variant of media literacy. Mixed findings may thus be explained by different features of the intervention as much as different features of the

[10] The content of the messages was related not to misinformation but to homosexuality, a contentious issue in Singapore.

context. Moreover, if the assumption is that people are not literate enough to detect misinformation, then media literacy interventions, unlike psychological ones, should only work after sufficient exposure.

Unusually, the evidence of effectiveness for media literacy interventions seems to be greater in the Global South than the Global North. However, the nature of the interventions is quite different. As is evident from Table 4, the most common media literacy intervention in the Global North is provided as a relatively short (one-page maximum) article that is read in text form by the respondent. By contrast, media literacy interventions in the Global South are mainly delivered by a several-minute video or in an interactive course lasting anywhere from one hour to eight weeks. Outcomes are more often measured immediately in the Global North and after some delay in the Global South (though sometimes both).

EVIDENCE ON MEDIA LITERACY FROM THE GLOBAL NORTH

There is only limited evidence that media literacy interventions reduce misinformed beliefs in the Global North. Additionally, most studies do not evaluate how effective these treatments are in the long term. The most promising evidence comes from Guess et al. (2020), a large, rigorous study testing Facebook's "Tips to Spot False News" and the only study in this intervention category to test for durability. This brief informational intervention improved discernment between true and false news headlines in the U.S. both immediately after exposure and to a lesser extent several weeks after the intervention. Another media literacy study, Domgaard and Park (2021), compared infographics against text-only tips. It found infographics were more effective in reducing beliefs about COVID-19 vaccinations, but the sample size of the study may be too small to draw robust conclusions.

Most studies from the Global North, however, find that media literacy interventions do not increase discernment and one even finds a negative effect. For example, a news literacy message did not affect the ability to recognize fake news on simulated Facebook (Vraga et al., 2021) or Twitter (Vraga et al., 2022). The negative finding comes from a study of the effect of lateral reading. It found that the intervention increased belief in the misinformation that participants were meant to be evaluating through searching online (Aslett et al., 2022). Because this negative effect is concentrated among respondents for whom searching turns up low-quality information, the authors conclude that interventions encouraging online searching should additionally teach individuals how to use proper search terms and identify quality news sources.

Table 4: Description of media literacy interventions

Global North	Global South
A box with strategies that readers can use to identify false or misleading stories that appear on their news feeds (Guess et al., 2020).	A box with strategies that readers can use to identify false or misleading stories that appear on their news feeds (Guess et al., 2020).
A half-page article detailing three tips to recognize misinformation (Hameleers, 2022).	A three-minute video about the perils of false news, including tips on how to identify them (Ali and Qazi, 2021).
A prompt encouraging users to utilize search engines (Aslett et al., 2022).	A four-minute informative video about online misinformation (Gottlieb et al., 2022).
A sponsored tweet warning about misinformation and encouraging critical news consumption (Vraga et al., 2022).	An hour-long, in-person learning module encouraging people to verify information along with provision of tools to do so (Badrinathan, 2021).

Global North	Global South
A pop-up on Facebook that presents a list of civic online reasoning techniques (e.g., lateral reading, click restraint) as tips to verify the information (Panizza et al., 2022).	A five-day course focused on teaching strategies to evaluate whether a post contains misinformation (Athey et al., 2022).
An infographic with an explanation of reverse image search and step-by-step instructions on how to use it (Qian et al., 2023).	A six-week program to create awareness of the problem of false news, with periodic assignments aimed at developing tools to deal with them (Apuke et al., 2022).
A one-page infographic with tips for finding false news (Domgaard and Park, 2021).	An eight-week course on digital and media literacy, including what fake news is and how to spot it (Apuke et al., 2023).
A 30-second video on Facebook with examples for how to spot misinformation (Vraga et al., 2021).	An eight-week program focused on the development of social media literacy skill training using visual multimedia package (Zhang et al., 2022).

Media literacy findings: Global North

Finding 1: There is only limited evidence that brief, scalable media literacy interventions can improve discernment (in the immediate and longer-term) in the Global North.

- One large, rigorous study finds that tips to spot fake news online improve discernment in the immediate and longer term.
- Most studies found that media literacy interventions do not positively affect discernment and one even found a negative effect.

Finding 2: Some studies find positive effects on alternative outcome measures but not on discernment itself.

Some studies find positive effects on alternative outcome measures but not on discernment. Textual media literacy interventions in the U.S. and the Netherlands reduced the perceived accuracy of untruthful statements about immigrants, but did not alter agreement with related anti-immigrant statement (Hameleers, 2022). When infographics were used in tandem with a provenance cue tool (see Figure 6), individuals reported higher intentions to use the tool in the future, but did not report decreased perceptions of credibility towards misinformed posts (Qian et al., 2023). In the United Kingdom, another intervention designed to teach individuals about lateral reading (i.e., fact-checking as one reads and encounters information) did not increase accuracy but did increase reported intentions to use the tool moving forward (Panizza et al., 2022).

EVIDENCE ON MEDIA LITERACY FROM THE GLOBAL SOUTH

Overall, media literacy interventions yield mixed results in the Global South. Variation in the effectiveness of interventions within and across studies suggests explanations for when we should expect these interventions to work best. This report identifies and discusses eight studies examining media literacy interventions in five Global South countries: Côte d'Ivoire, India, Kenya, Nigeria, and Pakistan.

Figure 6: Media literacy treatment in the U.S. from Qian et al. (2022)

Media literacy findings: Global South
Finding 1: Media literacy appears to improve discernment in the Global South most consistently among educated and tech-savvy participants. Effects appear to be weaker among populations with lower baseline education and literacy.
Finding 2: Interventions that are better tailored to the context or use non- standard approaches appear to work better than educational interventions that are purely information-based.
Finding 3: Interventions delivered over a longer period of time appear to work better than shorter-term interventions.

First, low baseline education and literacy rates appear to be a constraint on the effectiveness of media literacy interventions among representative samples of the population. The clearest evidence of this comes from Guess et al. (2020), a study in India that tested the same intervention on both types of samples — a highly educated online sample and a representative rural sample. The study, which assessed the impact of a campaign providing practical tips to spot misinformation, found a positive immediate effect among a highly educated online sample (though it was no longer measurable in a follow-up survey conducted weeks later). By contrast, the authors found no measurable results in a representative sample of largely rural areas in northern India, where education and social media usage were much lower. A more intensive media literacy intervention in India similarly failed to find an effect on a representative sample of the state of Bihar, one of the regions with lower levels of literacy rates in the country. In this context, Badrinathan (2021) found that an intensive, hour-long pedagogical intervention in which enumerators discussed many strategies designed to inculcate media literacy skills was ineffective. Finally, three studies found positive effects of a media literacy intervention among university students in Nigeria (Apuke et al., 2022, 2023; Zhang et al., 2022). Educational treatments, like the one depicted in Figure 7, require a relatively high level of baseline literacy and comprehension for the user to be able to effectively assimilate the information.

Second, interventions that are better tailored to the context or use non-standard approaches appear to work better than educational interventions that are purely informational. For example, a study in low- and middle-income areas of Lahore, Pakistan found that general video-based educational messages did not improve truth discernment, but they see a positive effect when the intervention is accompanied by personalized feedback based on the user's past engagement with false news (Ali and Qazi, 2021). A Kenyan study of approximately 9,000 English-speaking adults (which is thus a more educated sample than average) compares the effectiveness of a reasoning-based media literacy training against an emotion-based treatment and finds that the more standard media literacy intervention performed worse than the emotion-based treatment, although it still had a positive effect (Athey et al., 2022). In Côte d'Ivoire, a standard video-based digital literacy intervention providing skills to detect digital misinformation also finds null effects on discernment; however, an empathy-encouraging intervention in which an out-group individual described a life challenge they had faced was more promising (Gottlieb et al., 2022).

Figure 7: Media literacy treatment in Nigeria from Apuke et al. (2023)

Finally, more extensive interventions appear to be more effective than shorter- term interventions in this context. The successful educational course in Kenya was implemented over five days and the successful training in Nigeria was carried out over eight weeks — much longer than the unsuccessful one-hour workshop in India and several-minute video in Côte d'Ivoire. However, longer trainings are often more costly to provide and do not scale as easily.

In sum, this review of the evidence provides three suggestions for increasing the likelihood of success of media literacy interventions. First, implementers should consider the baseline education and literacy rates of the recipients before designing the study. A highly educated sample may benefit from short, information-based interventions similar to their counterparts in the Global North. But if the target

audience is comprised of a more representative sample of the population with varying literacy levels and exposure to digital media, then the evidence reviewed suggests two additional lessons. Interventions that are implemented over days or weeks may work better than brief interventions. Finally, interventions that are more tailored to the context by taking into account factors such as individual experiences, emotions, and social identities may be more effective than standard information-based interventions.

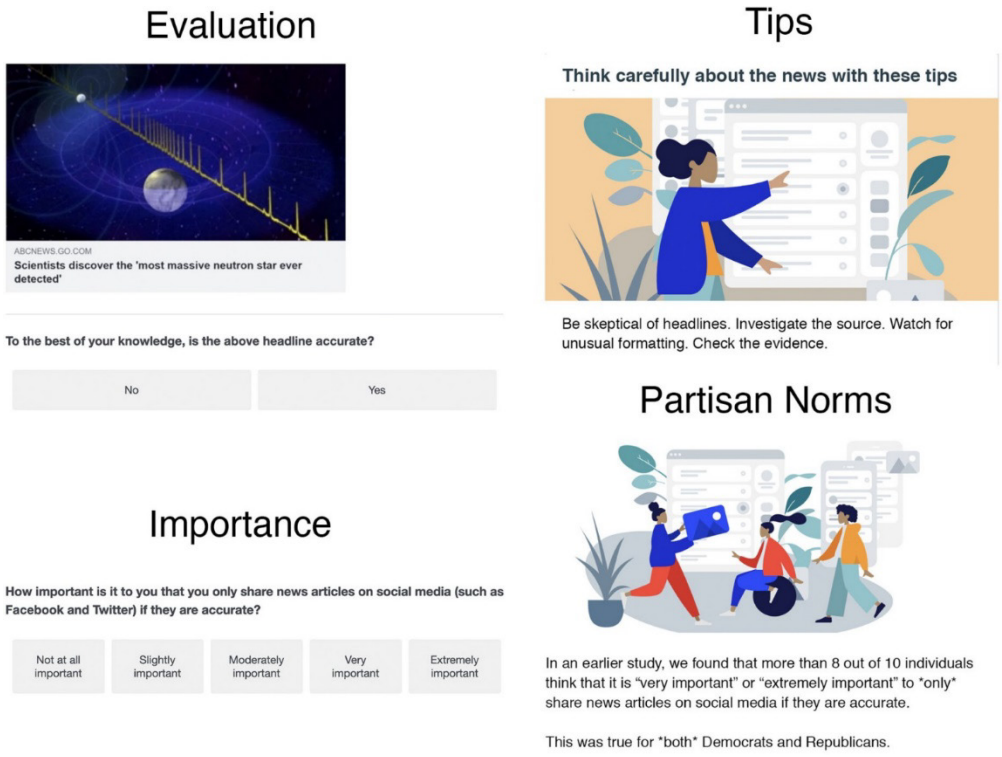
4.3 SOCIO-PSYCHOLOGICAL INTERVENTIONS

We define socio-psychological interventions as treatments that alter one's frame of mind or tap into an in-group and/or social identity. They do not provide direct information about a claim but instead prime the salience of a related concept in an individual's mind to try to reduce belief in and propensity to share misinformation. Accuracy prompts, friction and reflection prompts, and social/descriptive norms are the three intervention types that fall into this category. These studies too have largely been conducted in online survey experiments.

4.3.1 ACCURACY PROMPTS

Accuracy prompts seek to increase the salience of accuracy considerations in people's minds when they evaluate information. The idea is that people can better discern truth from falsehood when they are relatively more attentive to accuracy issues rather than factors unrelated to accuracy such as partisan alignment or group identities. Figure 8 displays a sample of the various types of accuracy prompts.

Figure 8: Accuracy prompt treatments from Epstein et al. (2021)



EVIDENCE ON ACCURACY PROMPTS FROM THE GLOBAL NORTH

Accuracy prompt findings: Global North

Finding 1: Accuracy prompts are generally effective in increasing discernment in the Global North, but effects vary in size.

Finding 2: There is variation in the mechanism increasing discernment: some interventions work by decreasing the sharing of false news while others work by increasing engagement with true content.

Accuracy prompt interventions are somewhat effective in increasing discernment in the Global North; however, the effects tend to be small and may not persist over time. Most studies in this area, which typically rely on survey experiments conducted among online samples in the U.S., seek to establish that people are better at discernment between true and false news when evaluating accuracy than when deciding whether to share truthful content, which suggests that individuals sometimes share misleading content because their attention is not focused on accuracy when deciding whether to share a piece of information (Epstein et al., 2021; Capraro and Celadin, 2022; Pennycook et al., 2020; Arechar et al., 2023; Bhardwaj et al., 2023). (We note, however, that a recent study does not find strong evidence of this assumption in other industrialized countries such as Italy and Spain (Arechar et al., 2023).)

The evidence generally suggests that accuracy prompts have a somewhat positive effect on discernment (Pennycook et al., 2020, 2021; Epstein et al., 2021; Roozenbeek et al., 2021; Jahanbakhsh et al., 2021; Capraro and Celadin, 2022; Guay et al., 2022; Epstein et al., 2023; Bhardwaj et al., 2023), but we note two important caveats. First, the magnitude of the effects significantly varies across studies. For instance, whereas Roozenbeek et al. (2021) and Rathje et al. (2023) find null or very small effects, Epstein et al. (2021), Epstein et al. (2023) and Guay et al. (2022) find quite substantial effects.

Second, uncertainty remains about the mechanism of the effect. While one set of studies find that accuracy prompts increase discernment by decreasing the sharing of false news (Epstein et al., 2021; Roozenbeek et al., 2021; Jahanbakhsh et al., 2021; Guay et al., 2022), others attribute the mechanism to increasing engagement with true content (Pennycook et al., 2020; Bhardwaj et al., 2023) and still others found both effects (Capraro and Celadin, 2022; Pennycook et al., 2021).

U.S. studies in this area further speak to heterogeneous treatment effects by partisanship and ideology. Among the studies that show positive effects from accuracy prompts, results are mixed. Some studies show that the effect exists for both liberal and conservative respondents (Epstein et al., 2021, 2023), while others suggest that effects are greater for people who identify with the Republican Party (Guay et al., 2022; Rathje et al., 2023). By contrast, a systematic meta-analysis shows that the effects are little to none among conservatives (Rathje et al., 2021), which is concerning given that some studies find lower levels of baseline of sharing discernment among this group (Jahanbakhsh et al., 2021; Pennycook et al., 2020; Grinberg et al., 2019).

EVIDENCE ON ACCURACY PROMPTS FROM THE GLOBAL SOUTH

Accuracy prompt findings: Global South

Finding 1: In the Global South, accuracy prompts are sometimes effective in increasing discernment and reducing sharing intentions, though effects are small.

In the Global South, evidence is mixed regarding the efficacy of accuracy prompts. We review three individual studies that test accuracy interventions across ten Global South countries: Argentina, Brazil, China, Egypt, India, Kenya, Nigeria, Philippines, Saudi Arabia, and South Africa.

Accuracy prompts reduced intent to share misinformation in both Kenya and Nigeria. In both countries, Offer-Westort et al. (2023) use an adaptive design on a sample Facebook users to test 40 factorial combinations of treatments, including various warning labels, media literacy tips and video training modules, and accuracy nudges. They find that their accuracy nudge, which was delivered via a chatbot on Facebook Messenger, decreased false news sharing by approximately 5% on average. Interestingly, the effects are substantial among participants who are not aligned with the governing party and have low levels of scientific knowledge. Athey et al. (2022) also test the efficacy of accuracy prompts (on their own and when paired with logic-based and emotions-based literacy courses) delivered over text messages. Among their sample of Kenyan participants, they find that accuracy prompts are effective at reducing sharing (about a 7% decrease) but less than the literacy courses. Moreover, adding accuracy prompts to the literacy courses did not increase their effects.

Other evidence on the effectiveness of accuracy interventions is more mixed. Of the 16 countries included in a study by Arechar et al. (2023), eight are located in the Global South. The findings vary by country — simply asking individuals to consider the accuracy of a non-COVID related headline substantially reduced intent to share false news in South Africa with smaller positive effects in India, the Philippines, Saudi Arabia, and China. However, no treatment effects were found in Argentina, Brazil, Egypt, Mexico, or Nigeria. They note that accuracy interventions "are unlikely to be helpful in countries where this disconnect is small (either because accuracy discernment is low or sharing discernment is already comparatively high), or for inaccurate claims that are widely believed" (15).

Which other factors could explain these mixed findings in the Global South? Unlike media literacy, accuracy treatments are quite similar across regions, so we cannot attribute it to differences in interventions. Most likely, as Arechar et al. (2023) suggests, people may simply find it more challenging to distinguish between true and false information in poorer countries with lower average education levels and less experience with online content,

Overall, accuracy prompts are somewhat effective on average, but may not be as effective as the interventions that were previously discussed. It is also unclear based on current evidence whether they are more effective among some groups compared to others. Finally, while these prompts are highly scalable and not dependent on the topic of misinformation, they do not seem to increase the efficacy of other, less scalable interventions.

4.3.2 FRICTION

Friction interventions encourage users to slow down and think before engaging with a given claim. These interventions seek to shift people from so-called System 1 processing (automatic, reliant on heuristics) to System 2 (more cognitively effortful and analytical), interrupting the process by which people may quickly believe or share misinformation because it is congenial or provokes strong emotions. The proposed mechanism sharply contrasts with theories of motivated reasoning, which typically posit that people who engage in effortful processing will seek additional arguments to protect their identity and reaffirm their pre-existing beliefs. In practice, these interventions include any treatment that encourages users to pause or, conversely, to discern the truth quickly. For an example of a friction intervention, see Figure 9 below. This review identifies three friction studies, all of which were tested in a U.S. context.

EVIDENCE ON FRICTION FROM THE GLOBAL NORTH

Friction findings: Global North

Finding 1: Friction may increase discernment and reduces sharing intentions for false news, but more experimental evidence is needed.

Friction interventions do tend to increase discernment between true and false news, although the available experimental evidence is scarce. Three studies in our database find a positive effect of frictions on discernment. In the U.S. and Canada, Sharevski et al. (2022) find that an interstitial cover, which obscures a tweet and states that it violated Twitter policy, reduced belief in false information. Bago et al. (2020) report that allowing individuals time to deliberate internally about the accuracy of true and false headlines decreased intentions to share false news.

Likewise, Fazio (2020) simply asks individuals to explain why a headline is true or false before indicating their likelihood to share, which reduces sharing intentions. However, Fazio (2020) also find that the treatment was less effective among those who viewed the headlines twice. If people are repeatedly engaging with the same piece of misinformation, friction interventions may be ineffective.

Figure 9: Friction intervention on Twitter from Sharevski et al. (2022)

In summary, these three studies demonstrate the initial promise of friction interventions and provide guidance for future testing in non-U.S. contexts. There are a variety of ways that friction prompts could be employed, and they have the benefit of being highly scalable because they are not specific to particular false claims. However, implementers should be aware that repeated exposure may render friction prompts ineffective in the long term.

4.3.3 SOCIAL NORM PROMPTS

Social norm interventions are typically messages from in-group members or other social media users aiming to discourage the sharing of misinformation by communicating behavioral expectations or standards. These interventions assume that, in some contexts, individuals could be motivated to engage with misinformation because it re-affirms their social identity, often in opposition to an out-group. We include several different definitions of prosocial or social norms-based interventions in this category. Some interventions fall under a sub-category of social or peer corrections in which other social media users remind or correct people about the perils of sharing misinformation. Other interventions (usually in the Global South) explicitly reference a social group when identifying the source of the message in order to prime in-group affect.

EVIDENCE ON SOCIAL NORMS FROM THE GLOBAL NORTH

Social norms findings: Global North

Finding 1: Social norm interventions are generally effective in the Global North.

Finding 2: In the U.S., there is mixed evidence regarding partisan source effects on corrections.

In the Global North, social norms interventions are generally effective at reducing both misinformed beliefs and sharing intentions. Studies typically define such interventions as any correction or reminder made by another social media user or the provision of information about other users that is meant to change perceived social norms. Under this definition, scholars have consistently found that social norms interventions reduce misperceptions and misinformation sharing in countries such as Australia (Ecker et al., 2022), Germany (Gimpel et al., 2021), Hungary (Orosz et al., 2023), the United States (Pretus et al., 2022; Andıand Akesson, 2021; Benegal and Scruggs, 2018; Berinsky, 2015), and the United Kingdom (Pretus et al., 2022).

Several of the studies considered investigate the influence of partisan identity on misinformation outcomes. One finds that corrective messages were more effective when they came from fellow party members compared to the opinion of general users, although this effect was only observed in the U.S. and not the less polarized context of the U.K. (Pretus et al., 2022). Two others demonstrate that partisan messages can be especially influential when the correction is not congruent with the expected partisan message (e.g., a Republican correcting climate misinformation or a Democrat correcting oil industry misinformation) (Berinsky, 2015; Benegal and Scruggs, 2018). Others, however, note that partisan source effects may be exaggerated. Clayton et al. (2019) find that neither Democrats nor Republicans are significantly swayed by partisan media sources when it came to news discernment (see Figure 10 below). In a similar vein, Chockalingam et al. (2021) show that, although out-partisan messages were viewed as less credible, the effects of a co-partisan correction and a standard correction not referencing a partisan source were not measurably different.

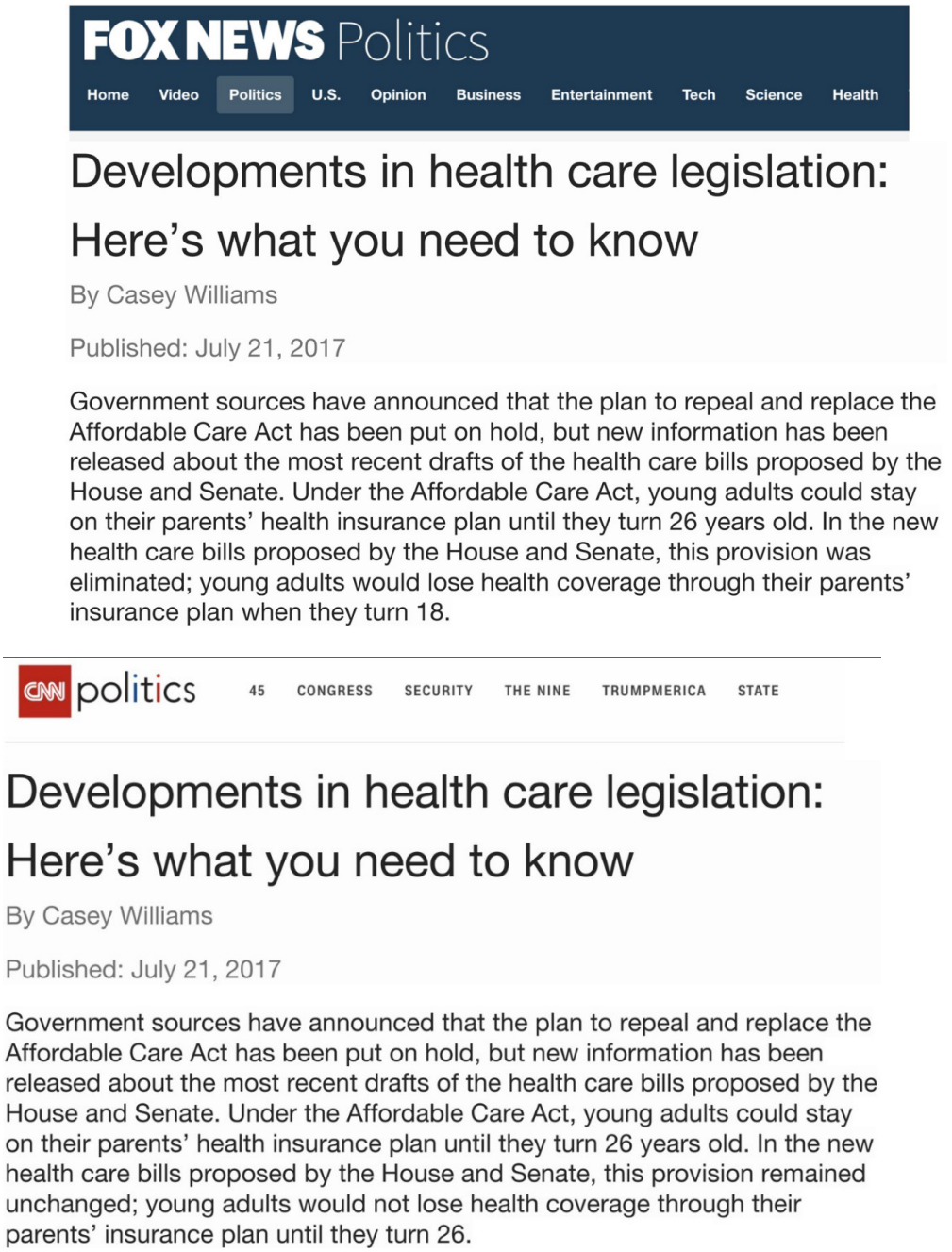
EVIDENCE ON SOCIAL NORMS FROM THE GLOBAL SOUTH

Social norms findings: Global South

Finding 1: A small set of studies from the Global South indicate that social norm interventions are effective, but more experimental evidence is needed.

A small set of studies finds social norms interventions are effective in the Global South as well. They evaluate such interventions in countries as different as Côte d'Ivoire (Gottlieb et al., 2022), India (Badrinathan and Chauchard, 2023; Pasquetto et al., 2020), Nigeria, and Pakistan (Pasquetto et al., 2020). Importantly, both Gottlieb et al. (2022) and Badrinathan and Chauchard (2023) measure the impact weeks after the treatment exposure, further strengthening the validity of the findings.

Figure 10: Experimental manipulation from Clayton et al. (2019)



A study conducted in India, Pakistan, and Nigeria sheds light on the type of social norms that are potentially more effective (Pasquetto et al., 2020). They found that voice-based corrections by a peer were more effective than text-based ones, and that corrections received from family members were re-shared more than ones received from acquaintances. However, they also found that this practice was unpleasant — sharing corrections was perceived as unusual and highly stressful and only very motivated individuals were willing to do so.

Some of these studies specifically seek to determine whether the observed effects on discernment are driven by a reduction in motivated reasoning. Specifically, they distinguish between effects on misinformation that relates to the individual's identity group and misinformation that is unaligned with identity. Motivated reasoning would potentially make people more vulnerable to misinformation belief in the former case than the latter because of the way that directional goals can override accuracy goals when people are evaluating information. Gottlieb et al. (2022) finds that an intervention intended to increase out-group empathy improves information discernment only when respondents are motivated to believe misinformation (because it affirms their identity), whereas Badrinathan and Chauchard (2023) finds no such a difference.

Even if the sum of evidence generally favors social norms interventions, these results should be interpreted with caution due to the multiple definitions of social norms across studies. If social norms interventions include any peer-to-peer message, then it is difficult to determine whether the effect is caused by social pressure or something else. In addition, the loose definition of an in-group also warrants caution. For instance, people may not consider a random Facebook user or a member of a WhatsApp group as part of their in-group. Thus, the success of a social norm intervention will depend on the context in which one intends to implement the intervention.

4.4 INSTITUTIONAL INTERVENTIONS

Any intervention that targets distributors of misinformation is included in the institutional category. We consider three interventions of this type: platform alterations, politician messaging, and journalist training. These have been tested using a mix of survey and field experiments.

4.4.1 PLATFORM ALTERATIONS

Platform alterations refer to changes in the interface or the algorithms used by platforms such as Facebook, Instagram, Twitter, YouTube, and WhatsApp that are intended to reduce the distribution or visibility of misleading or false content or engagement with it. This review considers studies that evaluate an alteration actually made by a platform as well as handful of examples of studies that simulate a platform environment and randomize some component of that environment. We caution that there is insufficient evidence to definitely evaluate the effectiveness of platform alterations as a group due to the limited evidence available to out- siders about internal testing at platforms and the many ways that such features can be implemented in practice (many of which overlap with the interventions described above). In other words, the effectiveness of these treatments depends on the specifics of the alteration, not on the mere fact of changing the platform architecture.

EVIDENCE ON PLATFORM ALTERATIONS FROM THE GLOBAL NORTH

Platform alteration findings: Global North

Finding 1: In the Global North, evidence is generally positive regarding the efficacy of various alterations to online platforms, but most evidence comes from simulating platform environments rather than changes to real platforms.

The evidence on platform alterations in the Global North, which comes entirely from the U.S., is generally positive. However, these studies face an important limitation. Due to the difficulty in obtaining

platform data or internal results, these studies almost exclusively test actual or proposed interventions in simulated plat- form environments among online survey participants rather than among real users on actual online survey platforms. Examples include studies that test credibility tags identifying posts as disputed or false, debunking misinformation through related stories, creating frictions in accessing misinformation, or de-platforming the leaders of hate organizations. Bode and Vraga (2015) and Sharevski et al. (2022) found a positive effect of debunking and friction-based interventions, respectively, Lees et al. (2022) found a positive effect of a credibility tag treatment but only among Democrats and independents, and Jennings and Stroud (2021) found modest evidence that credibility tag treatments reduce misperceptions across partisan divides. (These are example of studies of this type; other studies described above such as Clayton et al. 2020 also simulate elements of platform environments.)

We also note the effects of deactivating platforms entirely rather than altering their features. Deactivation studies encourage people to eliminate or substantially reduce usage of a social media platform. These interventions are blunt, as the treatment is the absence of exposure to all content from a platform, so it is difficult to disentangle which aspect of such a platform is responsible for any effect. In the U.S., Allcott et al. (2020) found that deactivating a Facebook account for four weeks reduced factual news knowledge and political polarization. In the Global South, two additional studies analyze the effectiveness of these interventions. In Bosnia and Herzegovina, Asimovic et al. (2021) found, contrary to their expectations, that Facebook deactivation somewhat decreased people's regard for ethnic out-groups. However, consistent with previous evidence, it also decreased knowledge of current events. In Brazil, Ventura et al. (2023) test the effect of deactivating WhatsApp in the weeks before the Brazilian presidential election in 2022, finding mixed results. On the one hand, deactivation reduced exposure to false news that circulated before the election; however, it did not cause significant changes in belief accuracy or political polarization.

Overall, our review of the evidence indicates that, when researchers simulate platform changes in experimental environments, their changes seem to be broadly effective. However, we note that the effectiveness of platform changes will depend on both the substance of the change and the way it is deployed online. Some platform alterations are more specific or didactic than others (e.g., debunking messages or corrective articles provided in response to misinformation are more intensive than a general friction or accuracy prompt). Moreover, the only evidence reviewed here from actual platform alterations focuses on deactivation, an extreme intervention which removes individuals and/or groups from a platform entirely. This type of intervention necessarily requires partnership with platforms in order to fully understand their impact. We encourage those with such connections to work together with academics to design experiments which will shed light on all that remains unknown about this type of institutional intervention.

4.4.2 POLITICIAN MESSAGING

These interventions seek to influence key suppliers of false information rather than consumers by warning politicians that their statements will be fact-checked and the results publicized. To our knowledge, only four studies have empirically tested the impact of these interventions, and none have done so in the Global South.

EVIDENCE ON POLITICIAN MESSAGING FROM THE GLOBAL NORTH

Politician messaging findings: Global North

Finding 1: Targeting politicians with messages that they are being monitored may aid in reducing the supply of misinformation, but more experimental studies are needed.

Additional evidence is necessary to ascertain how effectively misinformation can be addressed using elite interventions. Nyhan and Reifler (2015) were the first to experimentally evaluate a misinformation intervention targeting elites. In a field experiment with U.S. state legislators from nine states, the study tested whether politicians could be deterred from making misleading or misinformed claims. Those randomly selected for treatment received three letters over the course of three months leading up to the 2012 presidential election. Encouragingly, the legislators who received letters warning them that their statements would be fact- checked were less likely to have their accuracy criticized by either professional fact- checker PolitiFact or the media than the legislators in the study who did not receive letters. However, Ma et al. (2023) provide an update to Nyhan and Reifler (2015) by experimenting on state legislators in 2020 regarding Trump's impeachment. Treated politicians were sent three emails over the course of five weeks informing them of a partnership between the owner of several local television stations across the U.S., Hearst Communications, Inc., and FactCheck.org. The email included a clip of a recent media segment in which both a Republican and a Democratic legislator were fact-checked. They then observed tweets from the whole sample of state legislators to examine whether messaging politicians impacted the content they posted on Twitter. They do not find evidence that fact-checking deterred elites from sharing misinformation about Trump's impeachment on Twitter, which contrasts with the earlier positive findings in Nyhan and Reifler (2015).

Mattozzi et al. (2023) also investigate the effect of fact-checking on Italian MPs with a more intensive media campaign. Over ten weeks, the authors partnered with leading Italian fact-checker *Pagella Politica* to randomize fact-checks among the set of eligible MPs making false statements. These fact-checks were posted on the fact-checker's social media accounts, which tagged the treated politician's official Twitter account in the tweet advertising the fact-check. *Pagella Politica* also launched a video advertising the fact-check that was posted on websites and social media sites within two zip codes of the Italian parliament. They note that treated politicians reduce both the number of incorrect statements as well as the number of verifiable statements made relative to untreated politicians and that the effects persist for at least eight weeks.

Finally, Diermeier (2023) measures politician responsiveness to presentations of misinformation in Germany, especially among populist radical right parties. In a field experiment, every parliamentarian from the two legislative bodies in Germany (the Bundestag and the Bundesrat) received an email communication from what they believed to be one of their constituents on the topic of immigration, climate change, or unemployment. In reality, the emails were crafted by the research team and included a neutral but salient inquiry into whether a piece of misinformation regarding the topic at hand was true (see Figure 11). Interestingly, the Alternative für Deutschland (AfD) party, which emerged in 2013 and quickly gained support through campaigning on anti-immigration policies in 2021, had both the lowest response rate of all the parties and also the highest tolerance of the misinformation inquired about in the letter. Where only 5% of parliamentarians in other parties failed to contest the misinformed claim, 29% of AfD parliamentarians failed to contest the misinformation generally. Those statistics were even worse

for the migration misinformation letters; 40% of AfD elites failed to correct that misinformation compared to an average of 9% for the five other parties.

Taken altogether, the limited evidence presented in this report highlights both the promise of politician messaging as a way to combat misinformation as well as constraints to its effectiveness. While the content of this intervention may be scalable, conducting the actual fact-checking and/or other forms of monitoring likely requires a great deal of time and resources.

4.4.3 JOURNALISTIC INTERVENTIONS

The final institutional intervention aims to help journalists better communicate with the public about misinformation by identifying and debunking misinformation, disseminating messages indicating that media reports will be independently fact-checked, and/or applying weighting strategies to arguments. Of the three experimental studies we identified in this intervention category, only one was conducted in a non-American context. None are set in the Global South despite the interest in this type of intervention revealed by our expert survey (discussed below in Section 5). However, we do discuss a journalistic intervention tested in a Global South context that does not measure misinformation outcomes but may speak to the intervention's broad influence.

Figure 11: Artificial constituent letter from Diermeier (2023) (box added)

From: [alias]

Subject: Citizen's request: **Unemployment Corona /** Renewable energy **/** Migration

Dear Mrs. (Mr.) [name MP Bundestag/ Landtag],

my name is [alias]. I live in your constituency and am sending you this e-mail because I feel unsettled by **the current discussion about the crisis of the German economy during the Corona pandemic /** climate change **/** immigration to Germany. Your parliamentary activities as well as your party have helped a lot in the past and so I would like to ask for your help here as well.

Especially about the role of **the current situation of the labor market** / renewable energy **/** immigration in Germany there is a lot of different information. Since this topic plays a major role for me, I would now like to ask you for your personal assessment: How important do you consider the **current crisis of the German economy** / expansion of renewable energies in Germany **/** immigration for the German economy? And one piece of information I could not find despite research:

Is it true that **the unemployment in Germany is at 24 percent? /** only 35 percent of electricity consumption in Germany comes from renewable energies? **/** 48 percent of foreigners in Germany are unemployed?

Thank you very much for your help.

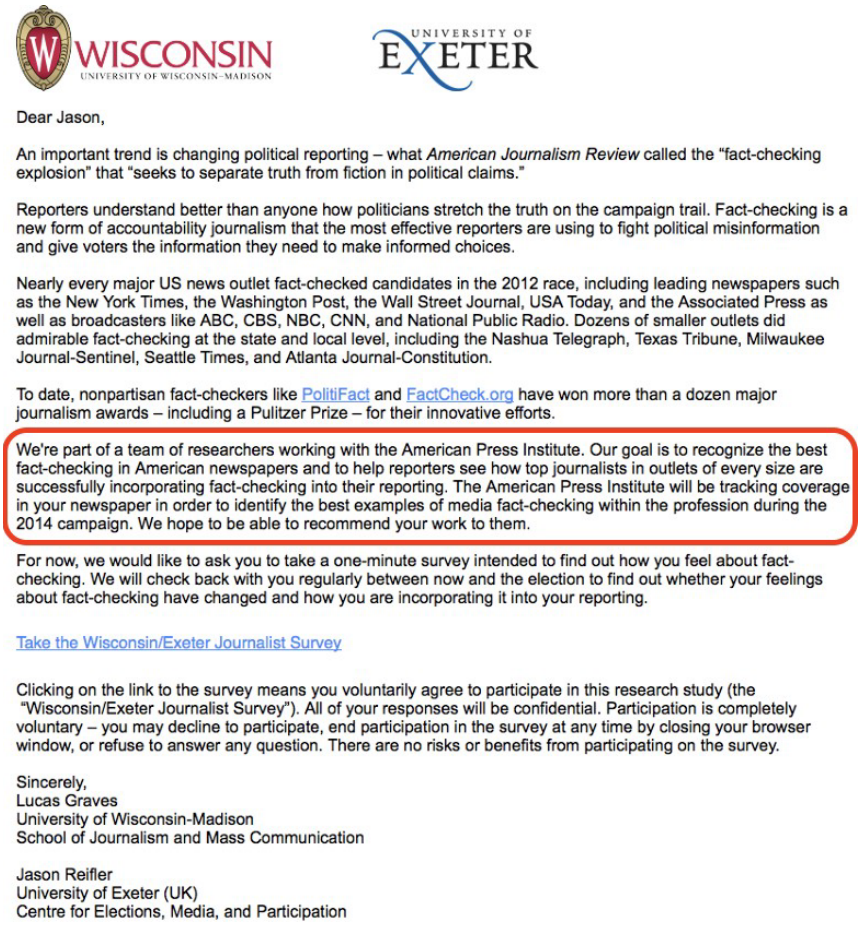
EVIDENCE ON JOURNALISTIC INTERVENTIONS FROM THE GLOBAL NORTH

Journalistic intervention findings: Global North

Finding 1: Journalistic interventions show initial promise, although more experimental studies are needed.

More experimental evidence is needed in both the Global North and Global South to better understand which journalistic interventions would be effective in combating misinformation. Two studies look exclusively at the American context. The first field experiment to be conducted among U.S. reporters tested potential motivations that might prompt journalists to engage more in fact-checking (Graves et al., 2016). Two different treatment letters were sent to the journalists at 82 newspapers —a supply-side treatment emphasizing fact-checking as professionally prestigious and reflecting the values of the profession (see Figure 12) and a demand-side treatment arguing that readers are hungry for fact-checking and that it attracts large audiences. Both treatment letters ended with statements that the journalists' coverage would be monitored by the American Press Institute in the hopes that they would be able to recommend the journalists' work to readers. The supply-side letter appealing to professional considerations was effective at increasing fact-checking coverage while the demand-side letter had no measurable effect.

Figure 12: Journalistic intervention from Graves et al. (2016) (box added)



WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

UNIVERSITY OF EXETER

Dear Jason,

An important trend is changing political reporting – what *American Journalism Review* called the "fact-checking explosion" that "seeks to separate truth from fiction in political claims."

Reporters understand better than anyone how politicians stretch the truth on the campaign trail. Fact-checking is a new form of accountability journalism that the most effective reporters are using to fight political misinformation and give voters the information they need to make informed choices.

Nearly every major US news outlet fact-checked candidates in the 2012 race, including leading newspapers such as the New York Times, the Washington Post, the Wall Street Journal, USA Today, and the Associated Press as well as broadcasters like ABC, CBS, NBC, CNN, and National Public Radio. Dozens of smaller outlets did admirable fact-checking at the state and local level, including the Nashua Telegraph, Texas Tribune, Milwaukee Journal-Sentinel, Seattle Times, and Atlanta Journal-Constitution.

To date, nonpartisan fact-checkers like PolitiFact and FactCheck.org have won more than a dozen major journalism awards – including a Pulitzer Prize – for their innovative efforts.

We're part of a team of researchers working with the American Press Institute. Our goal is to recognize the best fact-checking in American newspapers and to help reporters see how top journalists in outlets of every size are successfully incorporating fact-checking into their reporting. The American Press Institute will be tracking coverage in your newspaper in order to identify the best examples of media fact-checking within the profession during the 2014 campaign. We hope to be able to recommend your work to them.

For now, we would like to ask you to take a one-minute survey intended to find out how you feel about fact-checking. We will check back with you regularly between now and the election to find out whether your feelings about fact-checking have changed and how you are incorporating it into your reporting.

Take the Wisconsin/Exeter Journalist Survey

Clicking on the link to the survey means you voluntarily agree to participate in this research study (the "Wisconsin/Exeter Journalist Survey"). All of your responses will be confidential. Participation is completely voluntary – you may decline to participate, end participation in the survey at any time by closing your browser window, or refuse to answer any question. There are no risks or benefits from participating on the survey.

Sincerely,
Lucas Graves
University of Wisconsin-Madison
School of Journalism and Mass Communication

Jason Reifler
University of Exeter (UK)
Centre for Elections, Media, and Participation

Some studies don't directly intervene on journalists, but still speak to journalistic practices. Mena (2021) demonstrates that adding data visualizations to news articles does not increase their efficacy; all individuals who were presented with corrective messages reported decreased belief in misperceptions about immigrants and COVID-19 vaccines regardless of whether the article included a visual graph. Similarly, Schmid et al. (2020) investigates reporting strategies and their effects on misinformation belief among a sample of German students. Two weight-of- evidence strategies are tested: outnumbering and forewarning. Outnumbering is a corrective to the problem of false balance, which occurs when reporters or journalists give equal weight to both sides of a controversy despite the majority of evidence and consensus belonging to one side. Forewarning consists of informing the audience ahead of time that while both arguments have been given equal attention, it is not the case that both have an equal number of supporters. While outnumbering was not an effective strategy against climate science deniers, the forewarning treatment resulted in decreased belief in the science deniers' position. We identified one experimental study in the Global South that intervenes on journalists but does not explicitly measure misinformation outcomes. We discuss it here as it can have implications for affecting the provision of misinformation. Michelitch and Weghorst (2021) conducted a study with media studies students in Tanzania where participants engaged with Swahili materials designed "to generate practical skills in producing radio content and reduce bias in reporting around gender, age, and rural identities, attitudes and behaviors...training also included learning units on different news platforms, the role of internet and social media, how to edit and package a news story, and interviewing skills" (5). The training was implemented over the course of several months, making it one of the more intensive interventions. They do not find evidence that additional training increases journalistic knowledge, ethical and gender diversity, or student efficacy. They do note, however, that treated students expressed greater interest in covering topics related to women and rural interests relative to non-treated students. While Michelitch and Weghorst (2021) do not evaluate misinformation outcomes, they provide insight into challenges to developing training interventions in the Global South such as the importance of translated materials, consistency in intervention implementation across participants, the sample's technological literacy, and partnership with local organizations.

Overall, journalistic interventions show promise in combating misinformation. Providing journalists with professional incentives and/or tools to better report on topics subject to misinformation has the benefit of increased scalability because the effects can be applied to whatever false claims arise. However, much is still unknown about how individual journalists and organizations think about misinformation in the context of their professional role; potential heterogeneity in journalistic values could affect the efficacy of this intervention.

5. EXPERT SURVEY RESULTS

5.1 QUANTITATIVE ANALYSIS

Given the relative imbalance in the amount of evidence from Global North versus Global South countries, we complemented our evidence review with an expert survey. The main goal of the survey was to assess experts' expectations for which intervention types were most likely to work in Global South contexts.

We surveyed experts who work in both research and practice in areas related to misinformation and governance. We distributed the survey online via Qualtrics through existing practitioner and researcher email lists and networks of which the primary investigators and partners were already members. The survey was open throughout most of April 2023. It took respondents about 10 minutes to complete. We did not compensate them for their time.

To measure expert expectations on which interventions would be most likely to work in Global South contexts, we presented respondents with the following hypothetical allocation exercise: "imagine USAID has funds to allocate to programs aimed at combating misinformation in developing countries. Given this goal, how would you advise USAID to divide up [100 units of funding] across the 12 intervention types described below?" The "12 intervention types" referenced are the 11 included in this review plus technique rebuttal, which we originally classified as a separate intervention. After reviewing the literature, however, we decided to assign the studies that used it to the larger categories of debunking and inoculation as appropriate.

Findings from expert survey

Finding 1: Experts were most optimistic about the efficacy of educational and institutional interventions in the Global South (where evidence to date is most limited) and least optimistic about informational and socio- psychological interventions (where more evidence exists).

Finding 2: Experts on the Global North were generally most optimistic about platform alterations, while experts on the Global South were roughly equally optimistic about platform alterations, media literacy, and journalist training.

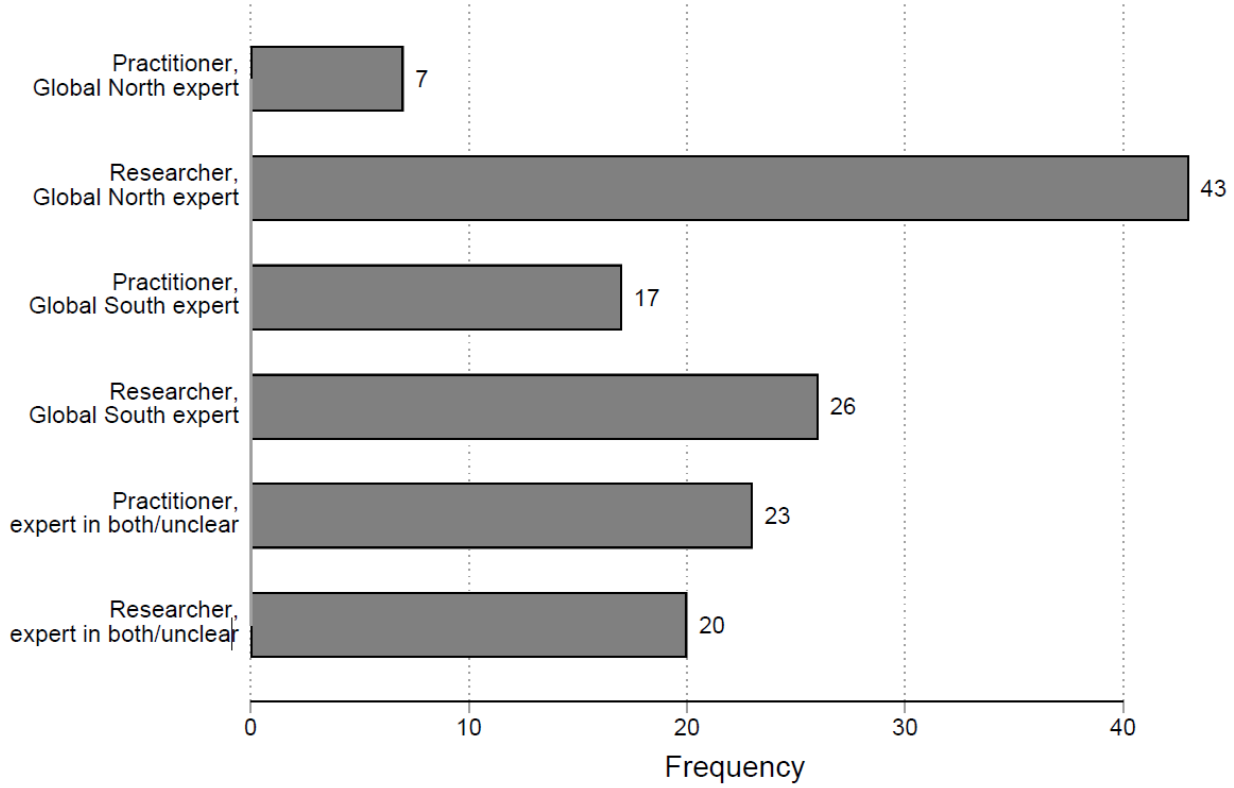
Finding 3: Experts generally believed that effective interventions from the Global North would be equally or less effective in the Global South; none thought they would be more effective.

A total of 138 experts responded to our main allocation exercise. We classify respondents along two dimensions. First, we code whether respondents are primarily researchers or practitioners/policymakers. Second, we code whether they are experts in the Global North, Global South, or both. Figure 13 displays the distribution of respondents along these two dimensions. (Note that region indicates the respondent's region of expertise, *not* their region of origin.) As expected, given the distribution channels, there are very few practitioner/policymakers who are strictly experts on the Global North, but many researchers who study the Global North.

Figure 14 displays the average amounts that respondents chose to allocate to each of the 12 intervention types, disaggregated by role and region of expertise. Re- searchers were most optimistic about platform

alterations, while practitioners were most optimistic about journalist training and media interventions. Respondents with expertise on the Global North (mainly researchers) were also most optimistic about platform alterations, whereas respondents with expertise on the Global South were most optimistic about media literacy, followed closely by journalist training and platform alterations.

Figure 13: Categories of respondents in the expert survey



Notably, respondents of all types and areas of regional expertise systematically preferred institutional and educational interventions over informational or socio- psychological ones. This finding is especially striking given that the evidence base is much stronger for informational and socio-psychological interventions than for educational and institutional ones, both in the Global North and Global South. In other words, the interventions about which experts were the most optimistic are also the ones about which we have the least evidence. Experts also tended to be most optimistic about interventions that are most difficult to study experimentally (e.g., platform alterations and journalist training).

As discussed in the review above, there are four types of interventions that have been shown to work in the Global North: inoculation (or prebunking), de- bunking, accuracy prompts, and frictions. However, relatively few studies have tested the effectiveness of these interventions in the Global South. We therefore asked respondents whether they expected these widely proven interventions to work better, worse, or equally well in the Global South relative to the Global North. For simplicity and because they rely on similar underlying logic, we combined prebunking and debunking into one category and accuracy prompts and friction into another for purposes of this exercise.

Figure 14: Mean allocations to each intervention type by respondent category

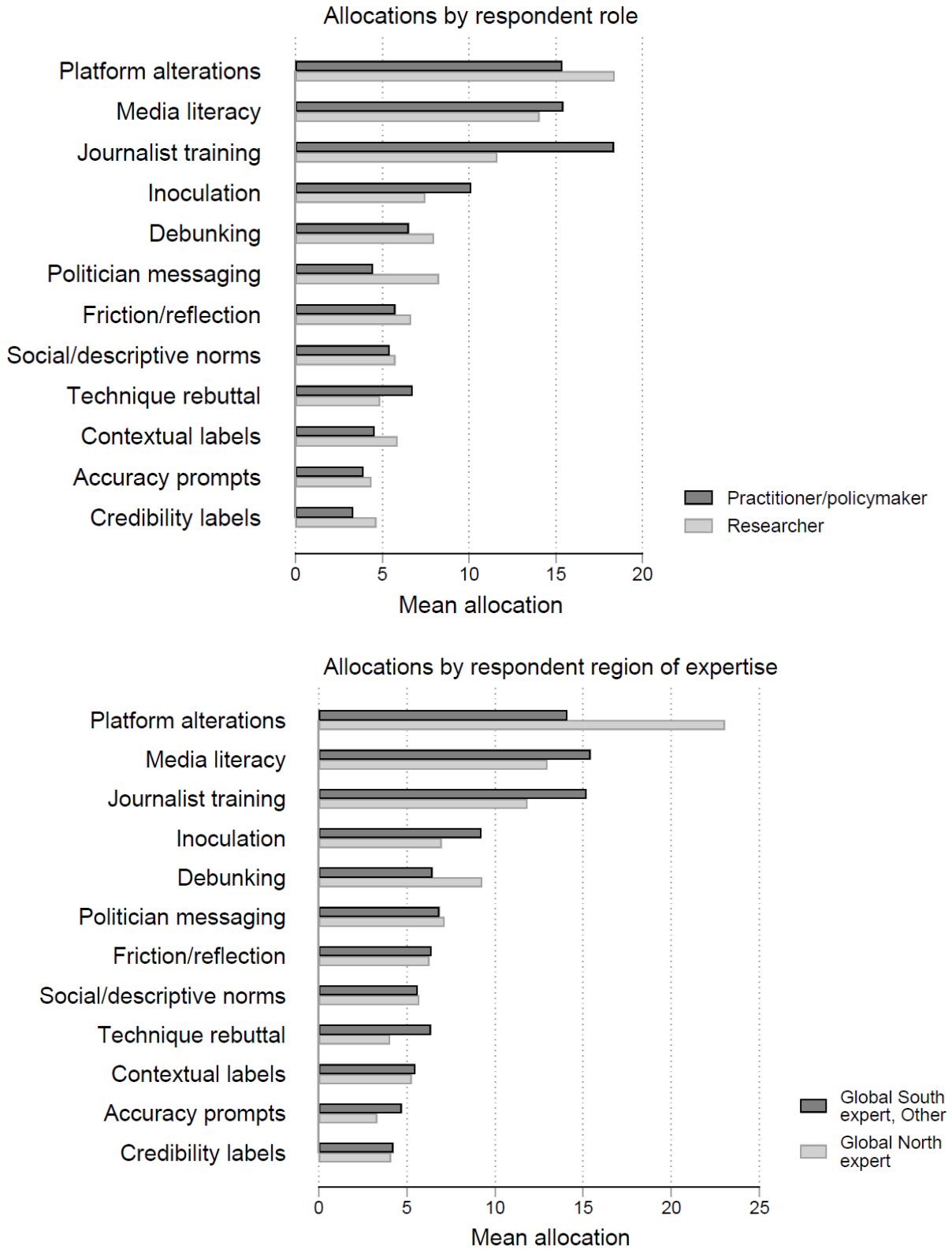


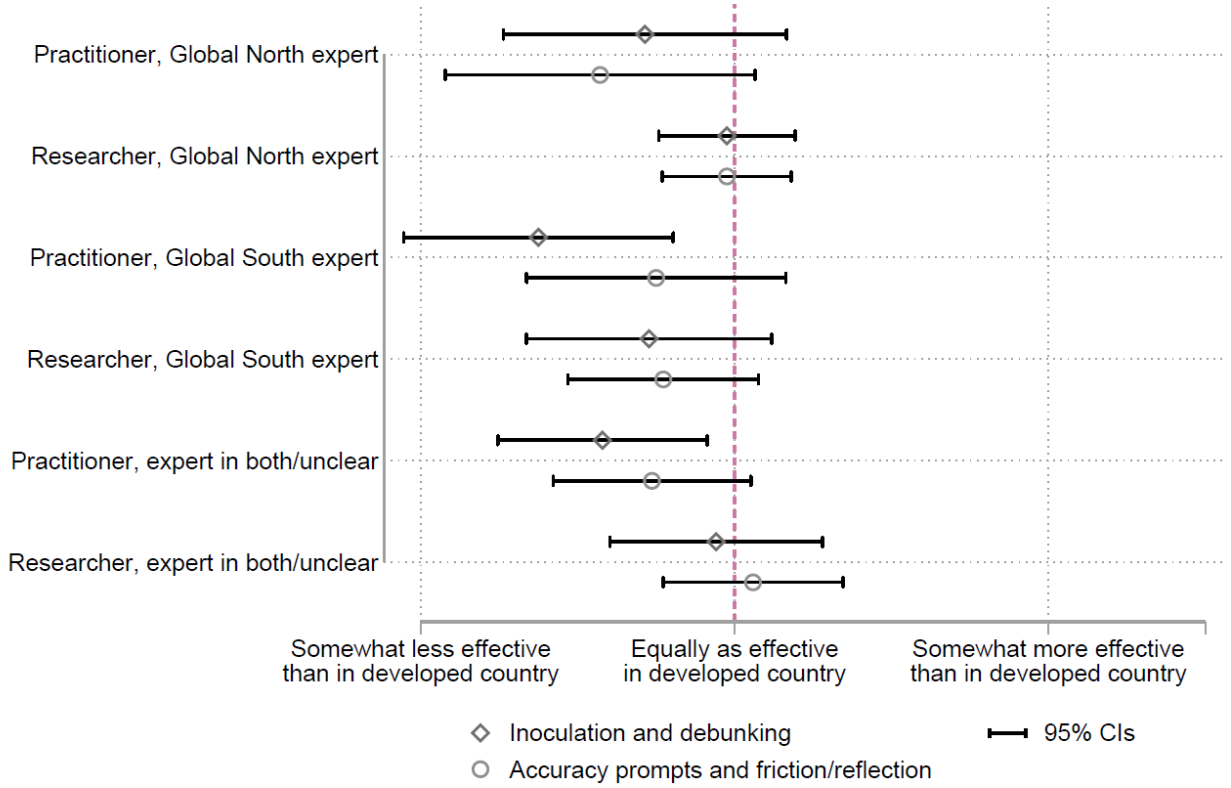
Figure 15 plots mean responses to this prompt among elites, again disaggregating by type and region of expertise. The bars denote 95% confidence intervals, which capture the amount of variation around the mean responses. (The wider the confidence interval, the more variation around the mean.) On average, researchers who focus on the Global North thought the interventions would be equally effective regardless of context. Practitioners and researchers focused on the Global South were more likely believe the interventions would be less effective in developing country settings. (We omit practitioners focused on the Global North from this figure because there are so few.) Notably, *no* category of experts, regardless of type or region of expertise, believed the interventions would be *more* effective in developing countries, which perhaps reflects the unique challenges of countering misinformation in the Global South.

5.2 QUALITATIVE ANALYSIS

In addition to the allocation exercise and closed-ended questions, we provided experts with the opportunity to elaborate on their responses in open-ended prompts. Specifically, we asked respondents to list "any interventions to combat misinformation that you think would work as well or better in developing countries than in developed ones. Please also briefly explain your reasoning." Of our 138 respondents, 102 offered responses to this prompt. We include a selection of responses below grouped by topic.

Many experts were optimistic and curious about interventions that target intentional or unintentional distributors of misinformation such as government actors and media professionals. As one North American researcher wrote, "working to train journalists should be equally or even more effective in developing countries, where journalists may have less initial training but are probably eager for more." Another researcher in North America agreed: "politician messaging might work better in developing countries due to politicians' dependence on foreign aid. Journalist training might also be more effective in developing countries due to the media having lower capacity in those settings." The capacity of journalists in the Global South was of concern to a third North American researcher who noted that "investments in journalist training/media interventions could show larger marginal effects in countries where journalism is severely under-resourced. In countries where press freedoms are restricted, however, this could increase risks to journalists, so the program would need to be very carefully designed."

Figure 15: Extent to which interventions would be *less* effective in a developing country context compared to a developed one



One concern is malicious actors. A practitioner in Eastern Europe noted that "in developing countries where misleading info is weaponised for political/personal/economic gain it is crucial to include forensic network analysis to detect coordinated behaviour and to map the malign actors, so they can be publicly exposed and (where appropriate/possible) deplatformed. Doing so destroys the digital infrastructure that has been built by mis/disinfo agents, driving up their operating costs, and demonetising the service economy that has grown to support toxic content campaigns." Specifically referencing politician propaganda, another Eastern European practitioner argued that "none of the interventions...are going to be tremendously effective in developing countries, especially in the contexts where the information environment is distorted by powerful propaganda campaigns coming from well-organized and sophisticated malicious actors such as Russia, China or domestic governments."

Relatedly, several experts referenced platform alterations and oversight in their responses. For instance, a North American researcher suggested that "platform alterations seem the only real answer, especially if we're considering language requirements + need to understand cultural context and nuance with emerging and rapidly spreading misinfo." One North American researcher raised the issue of detection, writing that "platform alterations are one of the most powerful tools, if not the single most powerful tool available for reducing engagement with misinformation. But for platforms to effectively reduce the reach of misinformation, they first need to effectively detect it. And platforms have devoted relatively few resources to misinformation detection in developing countries." This expert went on to argue that "investments in detecting and reducing the reach of misinformation will have larger marginal effects in developing countries than they would in developed countries, where platforms have already invested in the language and cultural competencies needed to comprehensively detect misinformation." A researcher in

Sub-Saharan Africa similarly suggested that more effective monitoring in the Global South would require that platforms hire "more people with contextual and local language knowledge."

Experts were also optimistic and curious about media literacy interventions in their open-ended responses. A Latin American practitioner argued that media literacy may operate by "encouraging citizens to make the assessments by them- selves," which may increase trust given that "sometimes they feel we are trying to convince them to be inclined to a certain side of a political aisle." Likewise, a Western European practitioner noted that "access to technology is increasing, but digital literacy and content literacy has not matched the speed at which access is increasing. Both digital natives and others need to be literate to analyze before sharing content." An East Asian researcher commented, "if the existing level of knowledge about misinformation and media literacy is low, I would expect media literacy, technique rebuttal, and journalist training/media interventions to be more effective in developing countries, due to a dose-response relationship (i.e., if existing knowledge is low, training is more beneficial)."

Notably, expert responses varied widely. Some respondents did not expect the effects of the interventions to differ by context:

> *"I don't expect meaningful heterogeneity by context of any of these approaches."*
>
> *"All of them. I have no particular reason to believe that any of them will work less well in developing countries."*
>
> *"I can't identify any of these as likely to be much more impactful in a developing context. In a very low literacy or internet penetration context, maybe some of the tactics focused on in-person misinformation (technique rebuttal, social norms) might be more effective relative to other strategies, but I couldn't say how meaningful that difference would likely be."*
>
> *"I think most interventions would work as well in developing countries as in developed ones."*
>
> *"It's hard to say unless we test them in developing countries. My hunch is that none works 'better' but some should work 'as well.'"*
>
> *"Given lower average education levels and lower baseline exposure to related interventions, I would expect all interventions that place relatively low cognitive demands and have low levels of abstraction to work better in developing countries."*

Others asserted that interventions should be highly tailored to fit local contexts:

> *"I don't that it makes sense to ask these questions without contextual information, and making broad conclusions would be misleading. I don't think this is a useful framing to use. What about asking under what conditions are these interventions most likely to be successful in the contexts in which they are most likely to be applied in these different countries?"*
>
> *"Disinformation is never about accuracy of specific facts, it is about the worldviews and geopolitics, disinformation actors are trying to manipulate emotions and trigger fears, so we will never be able to effectively combat it by rational arguments and techniques listed above. We need to understand what the values are, vision, perspectives and therefore narratives that strengthen resilience of the society."*

"I think it would depend on the local context—e.g., efforts to build trust in media or politics in areas with low pre-existing trust would likely yield better results over time as compared to regions with pre-existing higher levels of trust. Also, it is likely better to make small improvements in a range of areas (that affect a large number of people) than a big change in one area (that only limited number of people may be affected by)."

"I don't understand the prompt. You can't just compare all developed countries against all developing countries for misinformation. I'm not sure this is an even relevant dimension. You can talk about authoritarian regimes or countries facing war or some other dimension. But I would think that countering misinformation in India (at least pre-Modi) and France would be more similar than countering misinformation in India and China (both developing countries)."

6. HOW CONTEXT MODERATES EFFECTIVENESS OF INTERVEN- TIONS

The importance of context is a recurring theme in the literature on misinformation in the Global South. As discussed in Section 5 above, it was also a recurring theme in our expert survey. In some cases, the evidence base allows us to broadly compare the effectiveness of an intervention between the Global North and Global South. For the cases of debunking and inoculation interventions, there are enough studies from each context to allow us to draw some general conclusions. Both types of interventions appear to be effective regardless of context, though the moderators of effectiveness may vary. First, the duration of the intervention was found to improve intervention effectiveness in the Global South. Second, leveraging personal, political, or religious ties between the messenger and the information consumer was also found to increase the effectiveness of interventions there. Finally, differences in capacity (e.g., among fact-checkers, journalists, etc.) may constrain the effectiveness of interventions in the Global South.

In other cases, the way in which interventions are implemented and tested across the Global North and Global South preclude direct comparisons. In the case of media literacy, the standard intervention in the Global North is a brief text to be read by the information consumer. It is hard to compare this treatment with the videos and day- or week-long trainings that comprise media literacy interventions in the Global South. That said, examining variation in study effectiveness even within the Global South cases yields lessons that are quite similar to those discussed above. Intervening over longer periods of time again appears to improve effectiveness. And, adapting the intervention to the context by personalizing messages or leveraging identity or emotion appeared to improve effectiveness as well. One study finds differential effects across two samples within the same country: a brief media literacy intervention worked among a highly educated online sample but not a more representative rural sample. This finding suggests that baseline literacy might be a constraint to the effectiveness of such interventions.

But in some cases, there are no experimental studies in the Global South from which we can make inferences about context. In particular, for all the institutional interventions that intervene on information suppliers (journalists, politicians and platforms), we only found experimental studies from the Global North. Instead, we rely on intuitions from experts in our survey to inform our thinking about how context might moderate the effectiveness of these interventions.

As discussed in the previous section, experts noted that the lower capacity of journalists in the Global South could increase the effectiveness of journalist training interventions because there would be more

room for improvement. However, one of the findings from the media literacy studies suggests caution: it could be that baseline capacity is too low for relatively light-touch interventions to have any effect. Capacity to detect misinformation on platforms, especially in local languages, made experts skeptical about the effectiveness of platform alterations in Global South settings. Experts also warned that information suppliers in less democratic contexts would be less affected by training or monitoring interventions because they are less accountable and/or more vulnerable to malign influence.

To allow readers to make their own inferences about the role that context plays in the effectiveness of interventions, we include relevant contextual information about each study along several dimensions. For each country-year in the database of studies, we developed a complementary set of political and economic variables, obtained from the Varieties of Democracy Dataset (V-DEM), and the World Bank, respectively. This database will allow users to make targeted inquiries about what the evidence says in a specific type of context. For example, if a practitioner is planning an inoculation or debunking intervention in an autocracy, they can filter on the regime type variable and see what prior studies find in a less free political environment.

We sought to identify the most relevant contextual measures for which reliable cross-country data were available. The political variables provided in the database are regime type (closed autocracy, electoral autocracy, electoral democracy, liberal democracy), government social media monitoring, government social media censorship in practice, government Internet shutdown in practice, government social media alternatives, and media bias. The socioeconomic indicators included are GDP per capita and percentage of individuals using the internet. After selecting these variables, we calculated five-year moving averages for every country-year. For instance, if an intervention was conducted in Australia in 2020, we computed the average of the 2016–2020 period. Finally, we created terciles based on the global distribution of countries using the 5-year averages (except for regime type because it is ordinal).

The full database and the codebook are available here: https://www.democratic-erosion.com/briefs/misinformation-intervention-database/. The interactive table serves as a menu for possible interventions to implement that can be filtered on study-specific or context-specific indicators.

7. DISCUSSION AND RECOMMENDATIONS

7.1 PRACTICAL CONSIDERATIONS FOR IMPLEMENTATION

For policymakers and practitioners working to counter misinformation, *impact* (the initial efficacy of an intervention) is only one consideration in selecting and designing interventions. Additional considerations include *feasibility* (how easily an intervention can be implemented); *scalability* (how easily an intervention can be expanded to reach a large number of people); and *durability* (how long effects persist beyond the immediate post-intervention period). While it might be tempting to focus on interventions that score well on all four dimensions, such interventions rarely exist. Real-world interventions necessarily vary along these dimensions. Acknowledging these trade-offs can lead to more considered discussions of where to focus efforts and resources. Here we offer two practical examples of trade-offs that practitioners might face when deciding between intervention types.

DEBUNKING VS. ACCURACY PROMPTS

We first consider a hypothetical case in which a practitioner is partnering with an online media outlet in a Global South country that has the capacity to insert messages alongside news items. The practitioner is deciding between two types of messaging that have been proven to increase discernment between true and false news: debunking and accuracy prompts. Both have shown evidence of impact in the Global North and the Global South, but the effects of debunking tend to be larger and more durable. However, debunking messages typically only reduce false beliefs in the false claim that they target, reducing scalability. By contrast, accuracy prompts can affect beliefs irrespective of the topic. Furthermore, debunking requires media capacity to identify and refute misinformation unlike accuracy prompts, which are easily deployed across contexts. Thus, while debunking interventions promise potentially larger and more durable effects on a specific piece of misinformation, accuracy prompts offer greater feasibility and scalability. We note that the above discussion employs a broad definition of scalability comprising two distinct dimensions. An intervention can be viewed as scalable when it can easily be applied to a broader number of people (the most common definition). But we also might consider an intervention to be scalable when it is easily applicable to different phenomena or can be used to address a variety of problems or challenges. This is the definition employed above.

Deciding between interventions should be a function of the misinformation problem and the context. If a particular type of misinformation poses an important threat to health or security (e.g., misinformation about a particular group being immune to COVID-19), then debunking might be preferable. By contrast, if political opportunists are trying to stoke violence across ethnic groups by spreading multiple false images and narratives, accuracy prompts might be a better response.

MEDIA LITERACY VS. JOURNALIST TRAINING

Debunking and accuracy prompts require the partnership of a media outlet or platform if they are to be taken to scale. Let's consider a practitioner working in a context without such a partnership. In this case, the practitioner might consider tackling the misinformation problem via the demand side through a media literacy intervention or on the supply side by intervening directly on journalists. Both intervention types are seen as promising by experts who work in the Global South, but evidence of impact is either mixed or weak. In this case, deciding which intervention to employ might come down to feasibility and scalability: whether it is practical for the practitioner to reach large numbers of journalists or information consumers. This area is also one in which practitioners should consider partnering with researchers so that practitioners have more evidence on which to base their decisions in the future.

We summarize these two comparisons in Table 5, which represents the sort of exercise that practitioners might want to undertake when deciding amongst intervention types.

7.2 PROMISING AREAS FOR FUTURE RESEARCH

A large and growing body of empirical evidence has yielded important insights into the most effective interventions for curbing the spread of misinformation in the Global North and, to a lesser extent, the Global South. We conclude by suggesting some promising directions for future research.

Table 5: Qualitative assessments of interventions

Intervention	Impact	Feasibility	Scalability	Durability
Debunking	High	Medium	Low	High
Accuracy prompts	Medium	High	High	Low

Intervention	Impact	Feasibility	Scalability	Durability
Media literacy	Mixed evidence	Medium	Medium	Low
Journalist training	Weak evidence	Medium	Low	Medium

7.2.1 DESIGNING STUDIES THAT ALLOW FOR MORE DIRECT COMPARISONS BETWEEN GLOBAL NORTH AND GLOBAL SOUTH CONTEXTS

One goal of this report is to assess whether interventions that have proven effective at mitigating the spread of misinformation in the Global North might be equally effective in the Global South. As discussed in the introduction, this task is com- plicated by the fact that studies vary widely along many dimensions, including how participants are recruited; how interventions are designed and by whom, and how and when outcomes are measured. These differences make it hard to rule out the possibility that differing conclusions between studies conducted in the Global North and Global South are merely artifacts of differences in intervention and evaluation design.

The most straightforward way to address these ambiguities is to design studies that test the same interventions using comparable methods and samples in multiple countries simultaneously. While much more expensive and time-consuming than conducting a study in a single location, this approach allows for much more powerful inferences about the portability of interventions across contexts. Some researchers have already begun to take this approach, demonstrating its promise and feasibility. In the literature on debunking, for example, both Porter and Wood (2021) and Porter et al. (2023) administered fact-checks in multiple countries roughly at the same time, some in the Global South and others in the Global North. These studies offer a potentially replicable model for researchers wishing to consider a similar approach. (For another model, see the Evidence in Governance and Politics (EGAP) network's Metaketa Initiative: https://egap.org/our-work/the- metaketa-initiative/.)

7.2.2 EXPLORING WHETHER SOME TYPES OF MISINFORMATION ARE EASIER TO CURB THAN OTHERS

Some interventions may be more effective at controlling the spread of certain types of misinformation than others. Likewise, some types of misinformation may be less susceptible to correction than others. Misinformation about particular politicians or religious or ethnic groups, for example, often connects to deeply ingrained aspects of people's identities and may therefore be especially difficult to correct. Misinformation about public health likewise raises questions of life and death about which consumers may feel especially passionately. Most of the studies reviewed in this report focus on a particular topic or type of misinformation or even particular false or misleading claims. But a handful of researchers have begun incorporating multiple types of misinformation into their designs. Bowles et al. (2023), for

example, administer fact-checks covering a range of topics, including the economy, politics, race, COVID-19, crime, and other salient topics. Other studies could take a similar approach, generating insights that both researchers and practitioners could use to understand the extent to which interventions are broadly effective across different types of misinformation.

7.2.3 UNDERSTANDING THE ROLE OF SOCIAL IDENTITY IN EFFORTS TO COMBAT MISINFORMATION

Social identity plays a crucial role in the spread of misinformation. Consumers may be more likely to believe information they receive from sources they trust, such as those with whom they share political, racial, ethnic, or religious ties (Armand et al., 2021). Social identity may have equally powerful effects on efforts to combat misinformation. A growing number of informational and socio-psychological interventions have begun to explore this possibility. Pretus et al. (2022), for example, find that appeals to social norms in a highly polarized context like the

U.S. are more effective when the message originates with a co-partisan (e.g., a Democrat hearing from another Democrat). However, we caution that evidence on this phenomenon is mixed, as in-group source corrections may be no more effective than standard corrective messages (Clayton et al., 2019; Chockalingam et al., 2021). Nonetheless, social identity could be incorporated into a wide array of interventions. For example, to the extent that individuals' media diets reflect their political (or other) identities, media literacy interventions may prove more effective if they address the specific ways that misinformation spreads among the sources that are most popular among particular groups of consumers.

7.2.4 TESTING WHETHER INTERVENTIONS ARE MORE EFFECTIVE IN COMBINATION WITH ONE ANOTHER

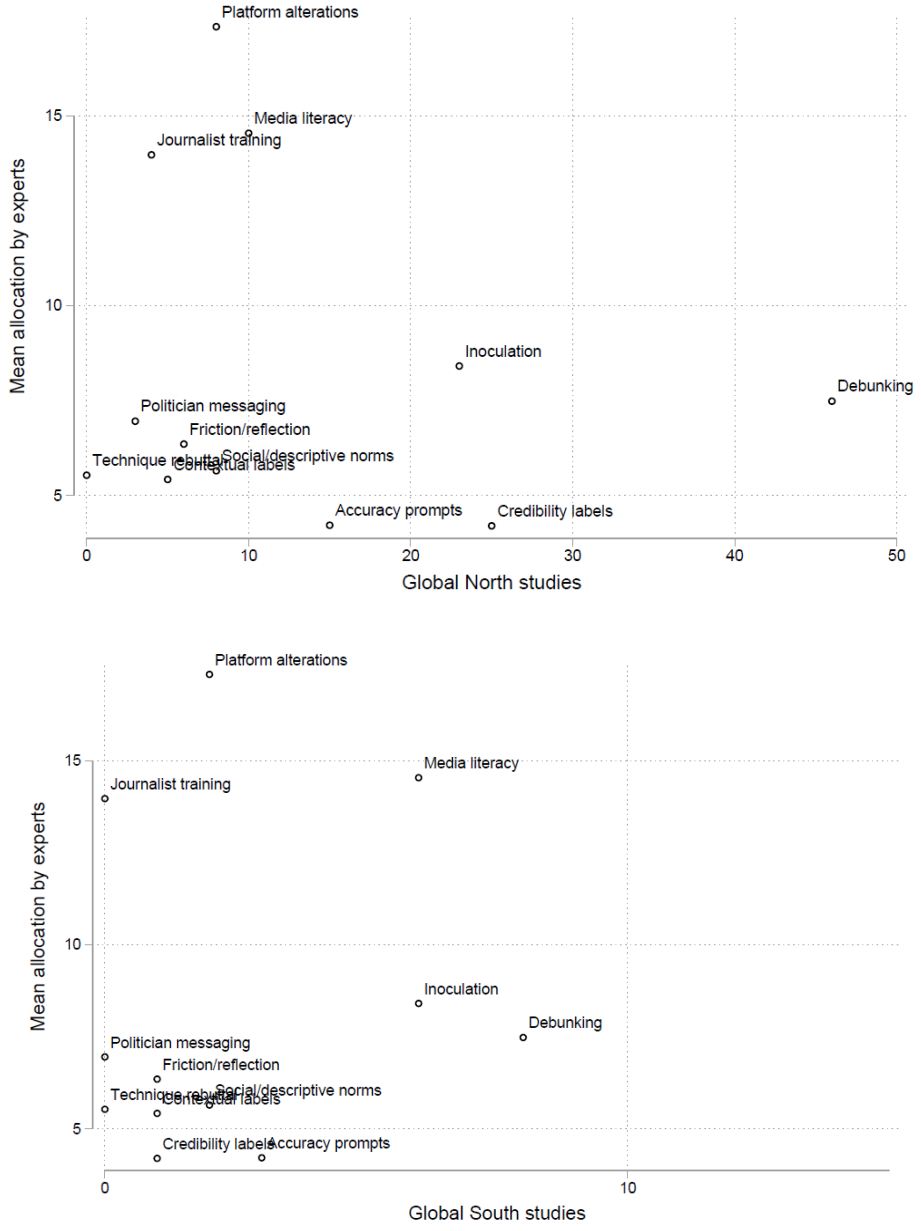
Most studies evaluate a single intervention or category of intervention in isolation. But some interventions may be more effective when combined with others and administered simultaneously or sequentially. (Conversely, some interventions may undermine one another — an equally important possibility to explore.) Only a handful of the studies covered in this review combine interventions in this way. Amazeen et al. (2022), for example, use prebunking and debunking in tandem to attempt to counter misinformation about the COVID-19 vaccines. Combinations of this sort are probably easiest to execute within rather than across our 11 categories of interventions (prebunking and debunking are both informational interventions). But combining interventions across categories may yield even more interesting insights. For example, is debunking more effective if it is combined with appeals to social norms? Is training for journalists more effective if it is combined with instruction in media literacy for the consumers of the stories journalists produce? Answering these questions may be especially valuable given that interventions are rarely implemented in isolation in the real world.

7.2.5 EXPANDING THE EVIDENCE BASE ON UNDERSTUDIED INTERVENTIONS

Finally and most obviously, some interventions have been much more extensively studied than others. For example, our search yielded 56 unique studies on debunking — the most of any intervention — but just four on politician messaging and three on journalist training. Likewise, the balance of the evidence is skewed much more heavily towards the Global North for some interventions than others. For example, approximately half of the 16 unique media literacy studies identified in our review were conducted in the Global North and half in the Global South. In contrast, of the 24 unique studies of credibility labels

and tags that we reviewed, only one tested the intervention in the Global South. These gaps are especially important given discrepancies between the robustness of the evidence on the one hand and the strength of experts' beliefs on the other. As we show in Figure 16, the three most popular interventions among experts — media literacy, journalist training, and platform alterations — are also among the least studied, with a total of 29 studies between them. This total is roughly equivalent to the number of unique studies on inoculation (25) or credibility labels (24) alone, and is less than half the number of unique studies on debunking (56), suggesting researchers should focus their efforts on interventions for which the evidence base is relatively weak and for which expert priors of expected effectiveness are particularly high. Such a focus strikes us as an especially important direction for future research.

Figure 16: Comparing expert evaluations with quantity of evidence



8. REFERENCES

Aird, M. J., U. K. Ecker, B. Swire, A. J. Berinsky, and S. Lewandowsky (2018). Does truth matter to voters? the effects of correcting political misinformation in an australian sample. *Royal Society Open Science 5*, 1–14.

Ali, A. and I. A. Qazi (2021). Countering Misinformation on Social Media Through Educational Interventions: Evidence from a Randomized Experiment in Pakistan. arXiv:2107.02775 [econ, q-fin].

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020, March). The Welfare Effects of Social Media. *American Economic Review 110*(3), 629–676.

Altay, S., B. Lyons, and A. Modirrousta-Galian (2023). Exposure to higher rates of false news erodes media trust and fuels skepticism in news judgment. PsyArXiv. Downloaded June 28, 2023 from https://psyarxiv.com/t9r43/.

Amazeen, M. A., A. Krishna, and R. Eschmann (2022). Cutting the bunk: Com- paring the solo and aggregate effects of prebunking and debunking covid-19 vaccine misinformation. *Science Communication 44*, 387–417.

Amazeen, M. A., E. Thorson, A. Muddiman, and L. Graves (2018). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism and Mass Communication Quarterly 95*, 28–48.

Amnesty International (2022). The social atrocity: Meta and the right to remedy for the rohingya. September 29, 2022. Downloaded June 27, 2023 from https://www.amnesty.org/en/documents/ASA16/5933/2022/en/.

Andı, S. and J. Akesson (2021). Nudging Away False News: Evidence from a Social Norms Experiment. *Digital Journalism 9*(1), 106–125.

APA News (2021). I/coast: Skirmishes ignited by fake news leave 10 wounded. May 20,2021. Downloaded June 27, 2023 from https://apanews.net/2021/05/20/icoast-ethnic-clashes-caused- by-fake-news-leave-10-people-wounded/.

Apuke, O. D., B. Omar, and E. Asude Tunca (2023). Literacy Concepts as an Intervention Strategy for Improving Fake News Knowledge, Detection Skills, and Curtailing the Tendency to Share Fake News in Nigeria. *Child & Youth Services 44*(1), 88–103.

Apuke, O. D., B. Omar, and E. A. Tunca (2022). Effect of Fake News Awareness as an Intervention Strategy for Motivating News Verification Behaviour Among Social Media Users in Nigeria: A Quasi-Experimental Research. *Journal of Asian and African Studies*.

Arechar, A. A., J. Allen, A. J. Berinsky, R. Cole, K. Garimella, A. Gully, J. G. Lu, R. M. Ross, Y. Zhang, G. Pennycook, and D. G. Rand (2023). Understanding and Combating Misinformation Across 16 Countries on Six Continents. PsyArXiv, https://psyarxiv.com/a9frz/.

Armand, A., K. K. Kameshwara, A. Bancalari, and B. Augsburg (2021). Countering misinformation with targeted messages: Experimental evidence using mobile phones.

Asimovic, N., J. Nagler, R. Bonneau, and J. A. Tucker (2021, June). Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences 118*(25), e2022819118. Publisher: Proceedings of the National Academy of Sciences.

Aslett, K., Z. Sanderson, W. Godel, N. Persily, J. Nagler, and J. A. Tucker (2022). Do Your Own Research: How Searching Online to Evaluate Misinformation Can Increase Its Perceived Veracity.

Athey, S., M. Cersosimo, K. Koutout, and Z. Li (2022). Emotion- versus Reasoning- based Drivers of Misinformation Sharing:.

Badrinathan, S. (2021). Educative interventions to combat misinformation: Evidence from a field experiment in India. *American Political Science Review 115*(4), 1325–1341.

Badrinathan, S. and S. Chauchard (2023). "I Don't Think That's True, Bro!" Social Corrections of Misinformation in India. *The International Journal of Press/Politics*.

Bago, B., D. G. Rand, and G. Pennycook (2020). Fake News, Fast and Slow: Deliberation Reduces Belief in False (but Not True) News Headlines. *Journal of Experimental Psychology: General 149*(8), 1608–1613.

Banas, J. A. and S. A. Rains (2010). A meta-analysis of research on inoculation theory. *Communication Monographs 77*, 281–311.

Bandeira, L., D. Barojan, R. Braga, J. L. Peñarredonda, and M. F. P. Argüello (2019). Disinformation in democracies: Strengthening digital resilience in Latin America. Atlantic Council, March 28, 2019. Downloaded June 27, 2023 from https://www.atlanticcouncil.org/in-depth-research-reports/report/disinformation-democracies-strengthening- digital-resilience-latin-america/.

Basol, M., J. Roozenbeek, M. Berriche, F. Uenal, W. P. McClanahan, and S. van der Linden (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against covid-19 misinformation. *Big Data and Society 8*, 1–18.

Benegal, S. D. and L. A. Scruggs (2018). Correcting misinformation about cli- mate change: the impact of partisanship in an experimental setting. *Climatic Change 148*, 61–80.

Bereskin, C. (2023). Understanding the efficacy of provenance interventions for tackling misinformation. Downloaded June 15, 2023 from https://osf.io/ 4a2fw.

Berinsky, A. J. (2015). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science 47*, 241–262.

Bertolotti, M. and P. Catellani (2023). Counterfactual thinking as a prebunking strategy to contrast misinformation on covid-19. *Journal of Experimental Social Psychology*.

Bhardwaj, V., C. Martel, and D. G. Rand (2023). Examining accuracy-prompt efficacy in combination with using colored borders to differentiate news and social content online. *Harvard Kennedy School Misinformation Review*.

Bode, L. and E. K. Vraga (2015). In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media: In Related News. *Journal of Communication 65*(4), 619–638.

Bode, L. and E. K. Vraga (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication 33*, 1131–1140.

Bowles, J., K. Croke, H. Larreguy, S. Liu, and J. Marshall (2023). Sustained exposure to fact-checks can inoculate citizens against misinformation. Un- published manuscript. Downloaded July 27, 2023 from https://scholar. harvard.edu/files/kcroke/files/wcw_submission.pdf.

Bowles, J., H. Larreguy, and S. Liu (2020). Countering misinformation via what- sapp: Preliminary evidence from the covid-19 pandemic in Zimbabwe. *PLoS ONE 15*, 1–11.

Brashier, N. M., G. Pennycook, A. J. B. E. Ó°Ä, D. G. Rand, and M. Levi (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences 118*(5), 1–3.

Buczel, M., P. D. Szyszka, A. Siwiak, M. Szpitalak, and R. Polczyk (2022). Vaccination against misinformation: The inoculation technique reduces the continued influence effect. *PLoS ONE 17*, e0267463.

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*.

Carey, J. M., V. Chi, D. J. Flynn, B. Nyhan, and T. Zeitzoff (2020). The effects of corrective information about disease epidemics and outbreaks: Evidence from zika and yellow fever in Brazil. *Science advances 6*, eaaw7449.

Celadin, T., V. Capraro, G. Pennycook, and D. G. Rand (2023). Displaying news source trustworthiness ratings reduces sharing intentions for false news posts. *Journal of Online Trust and Safety 1*, 1–19.

Chakroff, A. and R. Cole (2023). Provenance information reduces intent to share subtly manipulated images. Downloaded June 15, 2023 from https://psyarxiv.com/jptv4.

Chan, S., C. R. Jones, K. H. Jamieson, and D. Albarracín (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science 28*, 1531–1546.

Chockalingam, V., V. Wu, N. Berlinski, Z. Chandra, A. Hu, E. Jones, J. Kramer, X. S. Li, T. Monfre, Y. S. Ng, M. Sach, M. Smith-Lopez, S. Solomon, A. Sosanya, and B. Nyhan (2021). The limited effects of partisan and consensus messaging in correcting science misperceptions. *Research and Politics 8*(April–June), 1–9.

Clayton, K., S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan, M. Sandhu, R. Sang, R. Scholz-Bright, A. T. Welch, A. G. Wolff, A. Zhou, and B. Nyhan (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*, 1073–1095.

Clayton, K., J. Davis, K. Hinckley, and Y. Horiuchi (2019). Partisan motivated reasoning and misinformation in the media: Is news from ideologically uncongenial sources more suspicious? *Japanese Journal of Political Science 20*, 129–142.

Clayton, K., C. Finley, D. Flynn, M. Graves, and B. Nyhan (2021). Evaluating the effects of vaccine messaging on immunization intentions and behavior: Evidence from two randomized controlled trials in Vermont. *Vaccine 39*(40), 5909–5917.

Compton, J., S. van der Linden, J. Cook, and M. Basol (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass 15*, e12602.

Cook, J. (2016). Countering climate science denial and communicating scientific consensus. In *Oxford Research Encyclopedia of Climate Science*. Oxford University Press.

Cook, J., P. Ellerton, and D. Kinkead (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters 13*, 1–7.

Cook, J., S. Lewandowsky, and U. K. Ecker (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS ONE 12*, 1–21.

Dai, H., S. Saccardo, M. A. Han, L. Roh, N. Raja, S. Vangala, H. Modi, S. Pandya, M. Sloyan, and D. M. Croymans (2021). Behavioural nudges increase covid-19 vaccinations. *Nature 597*(7876), 404–409.

Dai, Y. N., W. Jia, L. Fu, M. Sun, and L. C. Jiang (2022). The effects of self- generated and other-generated ewom in inoculating against misinformation. *Telematics and Informatics 71*, 101835.

Diermeier, M. (2023). Tailoring the truth – evidence on parliamentarians' responsiveness and misinformation toleration from a field experiment. *European Political Science Review*.

Domgaard, S. and M. Park (2021). Combating misinformation: The effects of infographics in verifying false vaccine news. *Health Education Journal 80*, 974–986.

Ecker, U. K., L. H. Butler, and A. Hamby (2020). You don't have to tell a story! a registered report testing the effectiveness of narrative versus non-narrative misinformation corrections. *Cognitive Research: Principles and Implications 5*, 1–26.

Ecker, U. K., J. L. Hogan, and S. Lewandowsky (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition 6*, 185–192.

Ecker, U. K., S. Lewandowsky, and M. Chadwick (2020). Can corrections spread misinformation to new audiences? testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications 5*, 1–25.

Ecker, U. K., S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, and M. A. Amazeen (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology 1*(1), 13–29.

Ecker, U. K., J. A. Sanderson, P. McIlhiney, J. J. Rowsell, H. L. Quekett, G. D. Brown, and S. Lewandowsky (2022). Combining refutations and social norms increases belief change. *Quarterly Journal of Experimental Psychology 76*(6), 1275–1297.

Edgerly, S. and E. K. Vraga (2019). The Blue Check of Credibility: Does Ac- count Verification Matter When Evaluating News on Twitter. *Cyberpsychology, Behavior, and Social Networking 22*(4), 283–287.

Epstein, Z., A. J. Berinsky, R. Cole, A. Gully, G. Pennycook, and D. G. Rand (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review 2*(3), 1–12.

Epstein, Z., N. Sirlin, A. Arechar, G. Pennycook, and D. Rand (2023). The social media context interferes with truth discernment. *Science Advances 9*(9), 1–8.

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review 1*(2), 1–8.

Finance Center for South-South Cooperation (2023). Global south countries (group of 77 and China). Downloaded June 27, 2023 from http://www.fc-ssc.org/en/partnership_program/south_south_countries.

Freeze, M., M. Baumgartner, P. Bruno, J. R. Gunderson, J. Olin, M. Q. Ross, and J. Szafran (2021). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Political Behavior 43*, 1433–1465.

Gao, M., Z. Xiao, K. Karahalios, and W. T. Fu (2018). To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *Proceedings of the ACM on Human-Computer Interaction 2*, 1–16.

Garg, N., M. Yadav, R. Khera, E. Washington, E. Verhoogen, L. Boudreau, K. Zaremba, F. Grosset, O. Ahsan, and B. Seol (2022). Learning to resist misinformation: A field experiment in India.

Gimpel, H., S. Heger, C. Olenberger, and L. Utz (2021). The Effectiveness of Social Norms in Fighting Fake News on Social Media. *Journal of Management Information Systems 38*(1), 196–221.

Gottlieb, J., C. Adida, and R. Moussa (2022). Reducing Misinformation in a Polarized Context: Experimental Evidence from Côte d'Ivoire. Unpublished manuscript. Downloaded June 27, 2023 from https://osf.io/6x4wy/.

Grady, R. H., P. H. Ditto, and E. F. Loftus (2021). Nevertheless, partisanship persisted: fake news warnings help briefly, but bias returns with time. *Cognitive Research: Principles and Implications 6*, 1–16.

Graves, L., B. Nyhan, and J. Reifler (2016). Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. *Journal of Communication 66*, 102–138.

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science 363*(6425), 374–378. Publisher: American Association for the Advancement of Science.

Guay, B., A. Berinsky, G. Pennycook, and D. G. Rand (2022). Examining Partisan Asymmetries in Fake News Sharing and the Efficacy of Accuracy Prompt Interventions. Unpublished manuscript. Downloaded June 27, 2023 from osf.io/y762k.

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*, 15536–15545.

Guynn, J. (2016). Facebook unveils first serious effort to wipe out fake news. *USA Today*, December 15, 2016. Downloaded June 27, 2023 from https://www.usatoday.com/story/tech/news/2016/12/15/ facebook-taking-on-fake-news/95444334/.

Hameleers, M. (2022). Separating truth from lies: comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the us and Netherlands. *Information Communication and Society 25*, 110–126.

Hameleers, M., T. E. Powell, T. G. V. D. Meer, and L. Bos (2020). A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication 37*, 281–301.

Hameleers, M. and T. G. van der Meer (2020). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research 47*, 227–250.

Harjani, T., M.-S. Basol, J. Roozenbeek, and S. van der Linden (2023). Gamified inoculation against misinformation in india: A randomized control trial. *Journal of Trial and Error*.

Iyengar, A., P. Gupta, and N. Priya (2022). Inoculation against conspiracy theories: A consumer side approach to India's fake news problem. *Applied Cognitive Psychology 37*(2), 290–303.

Jackson, J., L. Kassa, and M. Townsend (2022). Facebook 'lets vigilantes in Ethiopia incite ethnic killing'. February 20, 2022. Downloaded June 27, 2023 from https://www.theguardian.com/technology/2022/feb/20/ facebook-lets-vigilantes-in-ethiopia-incite-ethnic-killing.

Jahanbakhsh, F., A. X. Zhang, A. J. Berinsky, G. Pennycook, D. G. Rand, and D. R. Karger (2021). Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. *Proceedings of the ACM on Human-Computer Interaction 5*(CSCW1), 1–42.

Jennings, J. and N. J. Stroud (2021). Asymmetric adjustment: Partisanship and correcting misinformation on facebook. *New Media and Society*.

Jiang, L. C., M. Sun, T. H. Chu, and S. C. Chia (2022). Inoculation works and health advocacy backfires: Building resistance to covid-19 vaccine misinformation in a low political trust context. *Frontiers in Psychology 13*, 1–11.

Johansson, P., F. Enock, S. Hale, B. Vidgen, C. Bereskin, H. Margetts, and J. Bright (2022). How can we combat online misinformation? a systematic overview of current interventions and their efficacy. arXiv. Downloaded June 28, 2023 from https://arxiv.org/abs/2212.11864.

Kasprak, A. (2016). Did 30,000 scientists declare climate change a hoax? Snopes, October 24, 2016. Downloaded June 27, 2023 from https://www.snopes. com/fact-check/30000-scientists-reject-climate-change/.

Kim, A., P. L. Moravec, and A. R. Dennis (2019). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems 36*, 931–968.

Kirchner, J. and C. Reuter (2020). Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-Computer Interaction 4*, 1–27.

Kozyreva, A., P. Lorenz-Spreen, S. Herzog, U. Ecker, S. Lewandowsky, and

R. Hertwig (2022). Toolbox of interventions against online misinformation and manipulation. PsyArXiv. Downloaded June 28, 2023 from https://psyarxiv. com/x8ejt/.

Lee, E.-J. and S. Y. Shin (2021). Debunking misinformation.

Lees, J., A. McCarter, and D. M. Sarno (2022). Twitter's disputed tags may be ineffective at reducing belief in fake news and only reduce intentions to share fake news among democrats and independents. *Journal of Online Trust and Safety 1*, 1–20.

Li, M., Q. Tay, M. J. Hurlstone, T. Kurz, and U. K. H. Ecker (2022). A comparison of prebunking and debunking interventions for implied versus explicit.

Ma, J., Y. Chen, H. Zhu, and Y. Gan (2023). Fighting covid-19 misinformation through an online game based on the inoculation theory: Analyzing the mediating effects of perceived threat and persuasion knowledge. *International Journal of Environmental Research and Public Health 20*, 1–18.

Maertens, R., J. Roozenbeek, M. Basol, and S. van der Linden (2021). Long- term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied 27*, 1–16.

Mattozzi, A., S. Nocito, and F. Sobbrio (2023). Fact-checking politicians.

McGuire, W. J. (1964). Inducing resistance to persuasion. some contemporary approaches. In C. Haaland and W. Kaelber (Eds.), *Self and Society. An Anthology of Readings*, pp. 192–230. Ginn Custom Publishing.

McPhedran, R., M. Ratajczak, and E. King (2022). Inoculation against misinformation in Australia: a replication study.

McPhedran, R., M. Ratajczak, M. Mawby, E. King, Y. Yang, and N. Gold (2023). Psychological inoculation protects against the social media infodemic. *Scientific reports 13*, 5780.

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy and Internet 12*, 165–183.

Mena, P. (2021). Reducing misperceptions through news stories with data visualization: The role of readers' prior knowledge and prior beliefs. *Journalism 24*(4), 729–748.

Meta (2019). Combatting misinformation on Instagram. December 16, 2019.

Downloaded June 27, 2023 from https://about.fb.com/news/2019/12/ combatting-misinformation-on-instagram/.

Michelitch, K. and K. Weghorst (2021). Impact evaluation of an intensive journalism training activity in Tanzania. USAID. Downloaded June 27, 2023 from https://pdf.usaid.gov/pdf_docs/PA00XM2D.pdf.

Modirrousta-Galian, A. and P. A. Higham (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis.

Nassetta, J. and K. Gross (2020). State media warning labels can counteract the effects of foreign disinformation. *Harvard Kennedy School Misinformation Review*.

Nekmat, E. (2020). Nudge effect of fact-check alerts: Source influence and media skepticism on sharing of news misinformation in social media. *Social Media and Society 6*, 1–14.

Nyhan, B. (2020). Facts and myths about misperceptions. *Journal of Economic Perspectives 34*(3), 220–236.

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences 118*(15), e1912440117.

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*, 939–960.

Nyhan, B. and J. Reifler (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior 32*, 303–330.

Nyhan, B. and J. Reifler (2015). The effect of fact-checking on elites: A field experiment on u.s. state legislators on jstor. *American Journal of Political Science 59*(3), 628–640.

Offer-Westort, M., L. R. Rosenzweig, and S. Athey (2023). Battling the Coronavirus Infodemic Among Social Media Users in Africa. arXiv:2212.13638 [cs, stat].

Orosz, G., B. Paskuj, L. Farago, and P. Kreko (2023). A prosocial fake news intervention with durable effects. *Scientific Reports 13*(1), 3958.

Panizza, F., P. Ronzani, C. Martini, S. Mattavelli, T. Morisseau, and M. Motterlini (2022). Lateral reading and monetary incentives to spot disinformation about science. *Scientific Reports 12*(1), 5678.

Pasquetto, I., B. Swire-Thompson, and M. A. Amazeen (2020). Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School Misinformation Review 1*(8), 1–14.

Pasquetto, I. V., E. Jahani, S. Atreja, and M. Baum (2022). Social debunking of misinformation on WhatsApp: The case for strong and in-group ties. *Proceedings of the ACM on Human-Computer Interaction 6*, 1–35.

Pennycook, G., A. Bear, E. T. Collins, and D. G. Rand (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science 66*, 4944–4957.

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, F. B., N. S. Bueno, F. Nunes, and N. Pavão (2022a). Fake news, fact checking, and partisanship: The resilience of rumors in the 2018 Brazilian elections. *Journal of Politics 84*, 2188–2201.

Pereira, F. B., N. S. Bueno, F. Nunes, and N. Pavão (2022b). Inoculation reduces misinformation: Experimental evidence from a multidimensional intervention in Brazil.

Porter, E., Y. R. Velez, and T. J. Wood (2023). Correcting covid-19 vaccine misinformation in 10 countries. *Royal Society Open Science 10*(1), 1–12.

Porter, E. and T. J. Wood (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, south Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences 118*(37), 1–7.

Pretus, C., A. Javeed, D. R. Hughes, K. Hackenburg, M. Tsakiris, O. Vilarroya, and J. J. Van Bavel (2022). The Misleading count: An identity-based intervention to counter partisan misinformation sharing. Unpublished manuscript. Downloaded June 27, 2023 from https://psyarxiv.com/7j26y/.

Qian, S., C. Shen, and J. Zhang (2022). Fighting cheapfakes: using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. *Journal of Computer-Mediated Communication 28*(1), zmac024.

Qian, S., C. Shen, and J. Zhang (2023). Fighting cheapfakes: Using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. *Journal of Computer-Mediated Communication 28*(1), 1–12.

Raj, S. (2022). In india, debunking fake news and running into the authorities. *New York Times*, September 22, 2022. Downloaded June 27, 2023 from https://www.nytimes.com/2022/09/22/world/asia/india- debunking-fake-news.html.

Rathje, S., J. Roozenbeek, C. Steenbuch, J. J. Van Bavel, and S. van der Linden (2021). Meta-analysis reveals that accuracy nudges have little to no effect for us conservatives: Regarding Pennycook et al. (2020). Psychological Science.

Rathje, S., J. Roozenbeek, J. J. Van Bavel, and S. van der Linden (2023). Accuracy and social motivations shape judgements of (mis)information. *Nature Human Behaviour 7*, 892–903.

Roozenbeek, J., A. L. J. Freeman, and S. van der Linden (2021). How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020). *Psychological Science 32*(7), 1169–1178.

Roozenbeek, J., R. Maertens, W. McClanahan, and S. van der Linden (2021). Disentangling item and testing effects in inoculation research on online misinformation: Solomon revisited. *Educational and Psychological Measurement 81*, 340–362.

Roozenbeek, J., C. S. Traberg, and S. V. D. Linden (2022). Technique-based inoculation against real-world misinformation. *Royal Society Open Science 9*, 1–13.

Roth, Y. and A. Achuthan (2020). Building rules in public: Our approach to synthetic & manipulated media. February 4, 2020. Downloaded June 27, 2023 from https://blog.twitter.com/en_us/topics/company/ 2020/new-approach-to-synthetic-and-manipulated-media.

Roth, Y. and N. Pickles (2020). Updating our approach to misleading information. May 11, 2020. Downloaded June 27, 2022 from https://blog.twitter.com/en_us/topics/product/2020/updating- our-approach-to-misleading-information.

Sangalang, A., Y. Ophir, and J. N. Cappella (2019). The potential for narrative correctives to combat misinformation. *Journal of Communication 69*, 298–319.

Schmid, P. and C. Betsch (2019). Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour 3*, 931–939.

Schmid, P., M. Schwarzer, and C. Betsch (2020). Weight-of-evidence strategies to mitigate the influence of messages of science denialism in public discussions. *Journal of Cognition 3*(1), 1–17.

Schmid-Petri, H. and M. Bürger (2022). The effect of misinformation and inoculation: Replication of an experiment on the effect of false experts in the context of climate change communication. *Public Understanding of Science 31*, 152–167.

Seo, H., A. Xiong, and D. Lee (2019). Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. Proceedings of the 10th ACM Conference on Web Science.

Sharevski, F., R. Alsaadi, P. Jachim, and E. Pieroni (2022). Misinformation warnings: Twitter's soft moderation effects on covid-19 vaccine belief echoes. *Computers and Security 114*, 102577.

Sherman, I. N., J. W. Stokes, and E. M. Redmiles (2021). Designing media provenance indicators to combat fake media.

Smith, C. N. and H. H. Seitz (2019). Correcting misinformation about neuroscience via social media. *Science Communication 41*, 790–819. Algorithmic corrections.

Tandoc, E. C., S. Rosenthal, J. Yeo, Z. Ong, T. Yang, S. Malik, M. Ou, Y. Zhou, J. Zheng, H. A. B. Mohamed, J. Tan, Z. X. Lau, J. Y. Lim, E. C. T. Jr, and S. O. R. J. Y. Z. O. T. Y. S. M. M. Y. Z. J. Z. H. A. B. M. J. T. Z. X. L. J. Y. Lim (2022). Moving forward against misinformation or stepping back? WhatsApp's forwarded tag as an electronically relayed information cue. *International Journal of Communication 16*, 1851–1868.

Ternovski, J., J. Kalla, and P. M. Aronow (2021). Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments.

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication 33*, 460–480.

Twitter (2023). Community notes: a collaborative way to add helpful context to tweets and keep people better informed. Downloaded June 27, 2023 from https://communitynotes.twitter.com/guide/en.

Vaidya, T., D. Votipka, M. L. Mazurek, and M. Sherr (2019). Does being verified make you more credible? Account verification's effect on tweet credibility.

Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Van Der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine 28*(3), 460–467.

van der Meer, T. G. and Y. Jin (2020). Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health Communication 35*, 560–575.

Ventura, T., R. Majumdar, J. Nagler, and J. A. Tucker (2023, May). WhatsApp Increases Exposure to False Rumors but has Limited Effects on Beliefs and Polarization: Evidence from a Multimedia-Constrained Deactivation.

Vivion, M., E. A. L. Sidi, C. Betsch, M. Dionne, E. Dubé, S. M. Driedger,

D. Gagnon, J. Graham, D. Greyson, D. Hamel, S. Lewandowsky, N. Mac- Donald, B. Malo, S. B. Meyer, P. Schmid, A. Steenbeek, S. van der Linden,

P. Verger, H. O. Witteman, and M. Yesilada (2022). Prebunking messaging to inoculate against covid-19 vaccine misinformation: an effective strategy for public health. *Journal of Communication in Healthcare 15*, 232–242.

Vraga, E., M. Tully, and L. Bode (2022). Assessing the relative merits of news literacy and corrections in responding to misinformation on twitter. *New Media and Society 24*, 2354–2371.

Vraga, E. K. and L. Bode (2017). Using expert sources to correct health misinformation in social media. *Science Communication 39*, 621–645.

Vraga, E. K. and L. Bode (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication 37*(1), 136–144.

Vraga, E. K., L. Bode, and M. Tully (2021). The effects of a news literacy video and real-time corrections to video misinformation related to sunscreen and skin cancer. *Health Communication 37*(13), 1622–1630.

Vraga, E. K., S. C. Kim, and J. Cook (2019). Testing logic-based and humor-based corrections for science, health, and political misinformation on social media. *Journal of Broadcasting and Electronic Media 63*, 393–414.

Vraga, E. K., S. C. Kim, J. Cook, and L. Bode (2020). Testing the effective- ness of correction placement and type on instagram. *International Journal of Press/Politics 25*, 632–652.

Walter, N., J. Cohen, R. L. Holbert, and Y. Morag (2020). Fact-checking: A meta- analysis of what works and for whom. *Political Communication 37*, 350–375.

Wang, Y. (2021). Debunking misinformation about genetically modified food safety on social media: Can heuristic cues mitigate biased assimilation? *Science Communication 43*, 460–485.

Winters, M., B. Oppenheim, P. Sengeh, M. B. Jalloh, N. Webber, S. A. Pratt,

B. Leigh, H. Molsted-Alvesson, Z. Zeebari, C. J. Sundberg, M. F. Jalloh, and

H. Nordenstedt (2021). Debunking highly prevalent health misinformation using audio dramas delivered by WhatsApp: Evidence from a randomised controlled trial in Sierra Leone. *BMJ Global Health 6*, e006954.

Wood, T. and E. Porter (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior 41*, 135–163.

Yaqub, W., O. Kakhidze, M. L. Brockman, N. Memon, and S. Patil (2020). Effects of credibility indicators on social media news sharing intent. Proceedings of the 2020 CHI conference on human factors in computing systems.

Zhang, L., T. O. Iyendo, O. D. Apuke, and C. V. Gever (2022). Experimenting the effect of using visual multimedia intervention to inculcate social media literacy skills to tackle fake news. *Journal of Information Science*.