American Political Science Review (2025) 1–24
doi:10.1017/S0003055424001394

Sustaining Exposure to Fact-Checks: Misinformation Discernment, Media Consumption, and Its Political Implications

JEREMY BOWLES *University College London, United Kingdom*
KEVIN CROKE *Harvard University, United States*
HORACIO LARREGUY *Instituto Tecnológico Autónomo de México, Mexico*
SHELLEY LIU *Duke University, United States*
JOHN MARSHALL *Columbia University, United States*

Exposure to misinformation can affect citizens' beliefs, political preferences, and compliance with government policies. However, little is known about how to durably reduce susceptibility to misinformation, particularly in the Global South. We evaluate an intervention in South Africa that encouraged individuals to consume biweekly fact-checks—as text messages or podcasts—via WhatsApp for six months. Sustained exposure to these fact-checks induced substantial internalization of fact-checked content, while increasing participants' ability to discern new political and health misinformation upon exposure—*especially when fact-check consumption was financially incentivized. Fact-checks that could be quickly consumed via short text messages or via podcasts with empathetic content were most effective. We find limited effects on news consumption choices or verification behavior, but still observe changes in political attitudes and COVID-19-related behaviors. These results demonstrate that sustained exposure to fact-checks can inoculate citizens against future misinformation, but highlight the difficulty of inducing broader behavioral changes relating to media usage.*

INTRODUCTION

Misinformation about politics, social issues, and public health is a growing concern. Such content—defined by its potential to generate misperceptions about the true state of the world—encourages beliefs and behaviors that are potentially harmful for both individuals and societies at large (Kuklinski et al. 2000; Nyhan 2020). Across the globe, the spread of misinformation on social media has been linked with citizens' distrust in politics and unwillingness to comply with government policies (Argote et al. 2021; Berlinski et al. 2023). By fueling ideological divides and increasing political polarization (Tucker et al. 2018), exposure to misinformation may have contributed to events such as the 2020 Capitol Hill riots and Brexit. In the Global South, where citizens are especially reliant on closed platforms like WhatsApp for information (Pereira et al. 2024), misinformation has been linked to lynchings and mass electoral mobilization in India and racial violence in South Africa (Allen 2021; Badrinathan 2021).

Interventions to limit the potential impact of misinformation most frequently engage in *debunking* or *prebunking* (Blair et al. 2024). Debunking facilitates learning through retroactively correcting specific pieces of misinformation, often by explaining why it is false and providing an alternative explanation (Nyhan and Reifler 2015). Prebunking, which is closely connected to inoculation theory (Cook, Lewandowsky, and Ecker 2017), entails warning individuals about the threat of misinformation through examples and preemptively providing knowledge to help them identify and resist it. Both prebunking (e.g., Guess et al. 2020; Pereira et al. 2024; Roozenbeek and Van der Linden 2019) and debunking (e.g., Henry, Zhuravskaya, and Guriev 2022; Nyhan et al. 2020; Wood and Porter 2019) have been shown to increase skepticism of misinformation.

Sustained exposure to fact-checks—one popular method of combating misinformation—leverages complementarities between debunking and prebunking. Fact-checking most obviously debunks by informing citizens about particular false (and true) claims. But, more generally, repeated engagement with fact-checks should also prebunk by increasing general awareness of misinformation, explaining the logic behind common forms of

Jeremy Bowles, Assistant Professor, Department of Political Science and School of Public Policy, University College London, United Kingdom, jeremy.bowles@ucl.ac.uk.
Kevin Croke, Assistant Professor, Harvard T.H. Chan School of Public Health, Harvard University, United States, kcroke@hsph.harvard.edu.
Horacio Larreguy, Associate Professor, Departments of Economics and Political Science, Instituto Tecnológico Autónomo de México, Mexico, horacio.larreguy@itam.mx.
Corresponding author: Shelley Liu, Assistant Professor, Sanford School of Public Policy, Duke University, United States, shelley.liu@duke.edu.
John Marshall, Associate Professor, Department of Political Science, Columbia University, United States, jm4401@columbia.edu.

Received: October 18, 2023; revised: May 06, 2024; accepted: November 13, 2024.

misinformation, and demonstrating information verification strategies. As a result, fact-checking potentially limits the harmful consequences of misinformation both by shaping citizens' discernment and verification of misinformation *upon exposure* and also by shaping media consumption choices, which affect the extent of exposure in the first place.

Despite these potential benefits, it is difficult to induce citizens to repeatedly consume fact-checks and internalize the lessons contained within them (Nyhan 2020; Walter et al. 2020). While fact-checked information can be effective when delivered in one-off forced consumption settings (e.g., Porter and Wood 2021), sustained consumption outside of the lab or online surveys competes against attention-grabbing content on traditional media, the internet, and now social media (e.g., Prior 2007). Furthermore, existing studies—which largely consist of testing single-shot efforts to combat misinformation—find that most effects attenuate significantly within a few weeks (Guess et al. 2020; Nyhan 2020; Porter and Wood 2021). The short-lived nature of these effects highlights the challenge of internalization, even conditional on information consumption (Zaller 1992), and points to the need to assess if continued fact-checking shapes political dispositions and compliance with the state beyond attitudes and behaviors closely connected to debunked misinformation.

To understand the consequences of sustained engagement with fact-checks in the field, we implemented a six-month field experiment via WhatsApp in South Africa, where misinformation about social, political, and health issues is rife (Servick 2015; Wasserman 2020). We partnered with Africa Check, the first fact-checking organization serving sub-Saharan Africa, to expose citizens to professionally produced fact-checks. Once every two weeks for six months, Africa Check delivered three fact-checks via WhatsApp messages to treated participants in our large rolling sample of social media users. These fact-checks dissected mostly false stories pertaining to politics, health, and other high-profile topics that were trending on social media in South Africa in the preceding weeks. To measure baseline demand for—as well as encourage the consumption of—the fact-checks, we cross-randomized whether treated participants received quizzes with financial incentives to correctly answer questions about the fact-checks or placebo quizzes containing questions about unrelated content.[1]

We further examine if, and how, citizens can be induced to engage with and internalize fact-checks by randomly varying how the fact-checks were disseminated to participants. Our four WhatsApp-based treatment conditions varied the appeal and cost of consuming the fact-checks as well as how empathetic the content was likely to be. First, imposing a low cost on consumers with competing time pressures, a simple text-based condition sent a single-sentence summary of each fact-check together with a weblink to additional information assessing each disputed claim. Second, the fact-checks were disseminated as a 6–8 minute podcast hosted by two narrators who fact-checked each claim and explained their verification process in a lively and conversational discussion that intended to generate engagement by making fact-checks entertaining.[2] Third, recognizing limits on time and attention span, we tested an abbreviated 4–6 minute podcast. Fourth, the full-length podcast was augmented with empathetic language emphasizing the narrators' understanding of how fear and concern for loved ones might lead individuals to be fooled by misinformation. These treatments build on literature relating to the challenges of ensuring citizens' attention to corrective information and news more generally (Baum 2002; Marshall 2023; Prior 2007), the effectiveness of "edutainment" in inducing behavioral change (Banerjee, La Ferrara, and Orozco-Olvera 2019; La Ferrara 2016), and the role of empathy in internalizing information (Gesser-Edelsburg et al. 2018; Gottlieb, Adida, and Moussa 2022; Kalla and Broockman 2020).

The results from our panel survey establish three core findings. First, we find that interest in fact-checks is difficult—but not impossible—to generate. While some participants engaged with the fact-checks in the absence of incentives, relatively small financial incentives generated substantially greater engagement with fact-checks during the intervention. Furthermore, sustained exposure to fact-checks significantly increased demand for future fact-checks absent the provision of incentives. This suggests that the intervention activated latent demand for entertaining fact-checks, as prior work encouraging citizens' access to novel news sources also finds (Chen and Yang 2019). These findings highlight the importance of attracting consumers for fact-checks to be effective at combating misinformation at scale.

Second, we demonstrate that sustained exposure to fact-checks helps to inoculate citizens against misinformation. Receiving any incentivized form of treatment persistently increased respondents' ability to discern truth from falsehood among a battery of political and public health stories we asked participants to assess, while increasing their skepticism towards prominent conspiracy theories—none of which were covered during the intervention. Importantly, greater discernment primarily reflected skepticism of false content, whereas confidence in true content was not systematically altered. Our results suggest that this may be driven by treated participants' greater attention to online content, increased understanding of what credible content looks like, and reduced trust in social media. Nevertheless, the treatments did not impact the amount of news that participants consumed from social and traditional media—and thus did little to change their risk of being

[1] The quizzes did not provide the correct answers, and thus provided incentives to consume fact-checks rather than constituting an additional information source. Moreover, since incentives were constant across conditions, we can isolate the effect of different conditions on information internalization upon consumption.

[2] Africa Check recorded these podcasts in partnership with the podcasting firm Volume. These podcasts were already a part of Africa Check's existing programming.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

exposed to misinformation—or their verification behavior. These findings suggest that sustained exposure to fact-checks primarily combats misinformation by increasing attention and skepticism upon exposure to false content, rather than by altering the type of content individuals consume in the first place.

Third, comparisons across treatment conditions indicate that the mode of dissemination matters. With respect to engagement, we find that less can be more: the quickly-consumable WhatsApp text message consistently produced larger effects on discernment than the more involved long and short podcasts. Furthermore, the text treatment shifted attitudes and reported behaviors relating to COVID-19 and government performance away from positions that could be fueled by misinformation: citizens became more likely to report complying with COVID-19 preventative behaviors recommended by the government and more favorable toward the current South African government. Only the empathetic version of the podcast increased discernment as much as the simple text messages, which suggests that edutainment is more effective when it includes emotive appeals to increase the resonance of corrective information with consumers.

Our study adds to the growing body of work studying interventions to hinder misinformation in the Global South (cf. Ali and Qazi 2023; Badrinathan 2021; Gottlieb, Adida, and Moussa 2022; Pereira et al. 2024; Porter and Wood 2021). A recent review of misinformation studies noted that more than 80% focused on contexts in the Global North, which "highlights the challenges of drawing conclusions about effective strategies for countering misinformation in the Global South" (Blair et al. 2024, 2). Our study's findings thus help to validate the benefits of fact-checking—which, through sustained exposure, essentially becomes media literacy—in settings where consumers have variable media literacy levels and face high data costs when independently validating information they find on social media platforms.

Our unusually sustained intervention in the field, along with the richness of our experimental research design, mean that our findings advance broader understanding of misinformation, how to combat it, and its political consequences in three key ways. First, we demonstrate that sustained exposure to fact-checks can not only debunk the specific misinformation addressed in the fact-checks but also prebunk new misinformation. The importance of repeated engagement helps to make sense of the mixed evidence on whether single-shot interventions can effectively prebunk misinformation (Maertens et al. 2021; Pereira et al. 2024; Roozenbeek and Van der Linden 2019; cf. Badrinathan 2021; Hameleers 2022). We also contribute to this literature by showing that interventions which encourage sustained exposure in a natural media consumption environment can be effective when citizens are motivated to consume fact-checks. By further measuring a broad array of outcomes, we establish that the enduring effects of our sustained intervention are largely driven by increasing citizens' capacity to discern content upon exposure, rather than by changing their media consumption habits. While the moderate effects we observe offer optimism for demand-side interventions, this finding simultaneously emphasizes the importance of complementary supply-side policies.

Second, our findings illuminate the theoretical mechanisms required for fact-checks to be effective at scale. In line with inventive studies seeking to "gamify" digital literacy lessons (Iyengar, Gupta, and Priya 2023; Maertens et al. 2021; Roozenbeek and Van der Linden 2019), we show that entertaining fact-checking podcasts can durably enhance citizens' discernment, and are most effective when delivered empathetically—as a growing literature suggests (Gesser-Edelsburg et al. 2018; Gottlieb, Adida, and Moussa 2022; Kalla and Broockman 2020; Williamson et al. 2021). However, we also show that "edutainment" is not the only pathway for stimulating engagement with, and internalization of, fact-checks. Indeed, short text messages that summarized fact-checks were at least as effective. Given the difficulty of engaging citizens in today's multi-platform media environment, interventions requiring little time commitment from citizens may be critical for conveying specific information and general lessons in the face of limited demand for fact-checks. This finding chimes with the importance of integrating brief accuracy nudges into social media platforms (e.g., Pennycook et al. 2021).

Third, this article addresses the important—but as yet understudied—question of whether misinformation shapes political attitudes and behaviors. While it is natural to believe that false beliefs might affect such outcomes, misinformed beliefs could instead reflect partisan cheerleading with limited political impact (Jerit and Zhao 2020). By demonstrating that text messages regularly conveying fact-checks both increased faith in the incumbent government and reported compliance with its policies, we show that reducing susceptibility to misinformation can have durable political consequences. Our results thus corroborate the perception that modern polities should be concerned about misinformation's potentially corrosive effects on state capacity and political accountability.

WHEN MIGHT FACT-CHECKING BE EFFECTIVE?

In much of the Global South, there are at least two important challenges to mitigating harmful exposure to misinformation. First, limited levels of digital literacy might amplify citizens' susceptibility to misinformation upon exposure (Badrinathan 2021; Guess et al. 2020; Offer-Westort, Rosenzweig, and Athey 2024). Second, high data costs restrict citizens' access to the broader internet and increase reliance on low-cost social media platforms such as WhatsApp (Bowles, Larreguy, and Liu 2020; Pereira et al. 2024). While platforms such as Facebook and Twitter can fact-check misinformation or warn users about flagged posts (Clayton et al. 2020), governments may lack the capacity or incentive to

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

encourage such interventions by platforms and these options are not possible for encrypted platforms like WhatsApp. Consequently, both citizens' overall exposure to misinformation, and the costs they face to verify it, are potentially high.

Research designed to mitigate the negative consequences of misinformation has predominantly focused on two types of interventions: corrective interventions (debunking) and preemptive interventions (prebunking). Corrective interventions, which debunk specific misconceptions and pieces of misinformation, are especially important for disproving prevalent or consequential claims of particular significance (Nyhan 2020). Conversely, prebunking efforts seek to "inoculate" people against specific claims, but also misinformation in general, by warning them about misinformation's existence and pre-emptively providing tools to identify and counteract it (Cook 2013; Martel, Pennycook, and Rand 2020).

Fact-checking is commonly associated with debunking, but may—with sustained exposure—combine both debunking and prebunking. While fact-checking interventions provide corrections about specific pieces of misinformation, fact-checkers often also explain the general steps taken to establish their conclusions. These explanations can inform consumers about the broader threat of misinformation and where it is most commonly found, demonstrate how misinformation can be debunked using reliable sources and how fact-checking techniques work, and explain general forms of faulty logic underlying specific false claims. Even without directly undergoing media literacy education, individuals may thus experience observational learning through repeated exposure to such messaging (Bandura 2001; Tewksbury, Weaver, and Maddex 2001; Zukin and Snyder 1984). Observational learning may even occur among individuals initially lacking motivation for active learning, where knowledge can still be acquired incidentally through passive consumption of information (Shehata et al. 2015; Stroud, Scacco, and Kim 2022). Fact-checking may therefore increase consumers' general awareness about how to avoid or spot misinformation and engage in critical thinking or fact-checking themselves.

Sustained exposure to fact-checks may then combat misinformation in two main ways. First, it could *reduce exposure* to misinformation. When consumers receive fact-checks consistently, they may become aware of the prevalence of misinformation, leading them to become more selective about what they read. As fact-checks also educate people about which types of sources are legitimate information providers, they may start consuming more reputable sources.

Second, sustained exposure to fact-checks can reduce misinformation's impact *upon exposure* by promoting internalization of the critical thinking skills they impart—which may require longer and more frequent exposure (Guess et al. 2020; Tully, Vraga, and Bode 2020). Thus, even if overall exposure to misinformation is not affected, internalization of the lessons from fact-checks may nevertheless ensure that individuals become more attentive to, discerning of, or likely to verify misinformation they encounter on social media or elsewhere; ideally, they would also become more trusting of truthful information. Sustained exposure may further enhance users' trust in fact-checking sources (Gentzkow, Wong, and Zhang Forthcoming), which may in turn increase internalization (Alt, Marshall, and Lassen 2016).

Although a number of studies experimentally demonstrate fact-checking's promise (see Blair et al. 2024), these studies also have important limitations (Flynn, Nyhan, and Reifler 2017; Walter et al. 2020). First, existing work primarily relies on one-shot interventions, often forcing participants to consume fact-checks in lab or survey environments. Outside these settings, however, citizens allocate their time across a wide array of activities and rarely choose to consume fact-checks. Various studies show that political news may only appeal to unusually-engaged individuals (Prior 2007) or when elections are upcoming (Marshall 2023), while relatively few people who visit untrustworthy websites get exposed to even one fact-check in the US (Guess, Brendan, and Reifler 2020)—let alone in the Global South, where mobile data is expensive. Corrective and preemptive interventions that work in the lab may then be of limited use in combating misinformation in the field if they cannot regularly capture the public's attention.

Second, consumption of fact-checks does not necessarily imply enduring internalization. Following Zaller (1992), people may read fact-checks and recall their content, but still fail to accept—and thus internalize—the information they receive or quickly move it to the back of their mind without repeated exposure. Indeed, some studies find evidence of motivated reasoning in response to counter-attitudinal information (Peterson and Iyengar 2021; Taber and Lodge 2006). Furthermore, existing research has tended to find short-term success only in combating the *specific* pieces of misinformation that fact-checks targeted, while failing to affect consumers' broader susceptibility or underlying attitudes or behaviors (Barrera et al. 2020; Carey et al. 2022; Hopkins, Sides, and Citrin 2019). Via either mechanism, limited internalization restricts fact-checking's potential benefits for media literacy.

Improving the Efficacy of Fact-Checks

Drawing from established theoretical frameworks, we consider how individuals might be encouraged to both consume and internalize fact-checks in the field.

Encouraging Engagement

Attracting consumers in a competitive media environment is likely to require reducing costs or increasing the benefits of consuming fact-checks. We first consider *reducing the time cost* of consumption. Competing against a flow of potentially more interesting or emotive content on social media and elsewhere, misinformation-correcting interventions that are quicker to digest for users might induce more consumption than interventions that take longer to ingest and understand. Given that internalization depends on initial consumption, easier-to-consume fact-checks may prove to be more effective at increasing audience reach and awareness.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

However, shorter interventions usually convey less information, so may have weaker effects on those exposed.

Another potential solution is to make fact-checking content *more appealing*. Following Bandura's (2001) application of social learning theory to mass media communication, using entertainment for educational objectives can promote attitudinal and behavioral change by increasing attention to favored behaviors, enhancing retention of modeled behaviors by making them more memorable, imparting the skills to reproduce their behaviors, and providing a strong motivation to carry out these behaviors. Prior research on "edutainment" shows that delivering information in entertaining and varied ways positively affects consumption, information recall, beliefs, and behaviors (e.g., Baum 2002; Baum and Jamison 2006; Kim 2023; La Ferrara 2016). For example, Banerjee, La Ferrara, and Orozco-Olvera (2019) find that exposure to television programming helped to increase awareness of HIV and health behaviors in Nigeria. Furthermore, Iyengar, Gupta, and Priya (2023), Maertens et al. (2021), and Roozenbeek and Van der Linden (2019) find that "gamified" media literacy training increased participants' likelihood of discerning between true and false tweets. Administering fact-checking interventions in more engaging ways might enhance users' demand for them.

Enhancing Internalization

The mode by which fact checks are delivered also has the potential to shape citizens' internalization. Within the literature, there is little consensus on the most effective modes of fact-checking, both when considering the *level of detail* or *tone of delivery* needed to inhibit susceptibility to misinformation. With respect to detail, lengthier fact-checks might appear more credible (Chan et al. 2017) and increase information retention (Lewandowsky et al. 2012); they also allow the fact-checking organization to provide more tips on how to spot, and verify, potential misinformation. Moreover, more detailed fact-checks may increase information retention and thereby boost media literacy (Lewandowsky et al. 2012). On the other hand, shorter messages may be less taxing on readers' attention, leading to greater engagement and, in turn, greater internalization (Pennycook et al. 2021). By reducing nuance, shorter and simpler interventions' concise takeaways might increase consumers' acceptance and recall of the fact-checked information (Walter et al. 2020).

Considering the tone of delivery, prior work points to the potential role of empathy in promoting internalization. An expanding body of work highlights the role of emotions in increasing susceptibility to misinformation (Martel, Pennycook, and Rand 2020). Thus, interventions that promote emotional engagement and empathy could induce sustained internalization (Gesser-Edelsburg et al. 2018). More generally, Kalla and Broockman (2020) show that empathetic narratives durably decreased outgroup exclusion, while Williamson et al. (2021) find that shared experiences, which induce empathy, increased support for immigrants.

However, the role of tone remains contested in the context of fact-checking. Bode, Vraga, and Tully (2020) find no improvement using either uncivil or affirmational tones in comparison to neutral-toned misinformation corrections. Martel, Mosleh, and Rand (2021) similarly find no impacts of polite corrective messages on the likelihood of engagement on social media or internalization of the misinformation correction. Since the inclusion of empathetic narratives is likely to increase the length of the fact-checks, trading detail for tone of delivery could reduce the effectiveness of empathetic fact-checks.

Hypotheses

Together, we anticipate that sustained exposure to fact-checking ought to combine aspects of both debunking and prebunking for misinformation correction. We next summarize our hypotheses, which were registered in our pre-analysis plan and are enumerated alongside a preview of our findings in Table 1.

Our hypotheses relate to four main groups of outcomes: engagement with fact-checks, discernment of content on social media, engagement with content on social media, and the political consequences of any changes in how individuals engage with the content they encounter. First, hypothesis H1 expected that easing access to fact-checks would increase exposure to, and knowledge of, the facts that were covered by the treatment deliveries. In a media environment with many available options, ensuring engagement with such content is not a given and may require further incentives. Second, we hypothesized that sustained exposure to corrective information would inoculate individuals against misinformation by making them more attentive to the veracity of the content they encounter (H2), more aware of misinformation on social media and less trusting of social media content (H3), and ultimately better able to discern false from true information (H4). Third, in addition to altering how citizens respond to content upon exposure, we further anticipated that repeated exposure to fact-checks would cause individuals to consume and share less content from social media (H5) and more actively use verification techniques (H6). Finally, to the extent that misinformation typically focuses on salient false claims about politics or public policy, sustained exposure to fact-checks might then increase compliance with government policies (H7) and improve perceptions of government performance (H8).

Beyond the effects of sustained exposure to fact-checks in general, understanding *how* to effectively increase organic consumption and internalization is more theoretically ambiguous. Indeed, simpler interventions might promote consumption while undermining the broader benefits from internalization, while more engaging modes enhance internalization but require more costly consumption decisions by citizens. Our pre-specified expectation relating to this trade-off was that interventions leveraging "edutainment" or more empathetic content would be more effective at enhancing internalization at the potential cost of lower

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

TABLE 1. Preregistered Hypotheses and Findings

Hypothesis in pre-analysis plan	Preregistered direction	Finding
Pooled effects		
H1: Exposure to, and knowledge about, information covered by treatment (Figure 4a, 4b)	+, +	+, +
H2: Attention to veracity of content on social media (Figure 6b)	+	+
H3: Perceived extent of true information on social media and trust in social media content (Figure 6c)	–	–
H4: Capacity to identify misinformation based on its characteristics (Figure 5a, 5b)	+	+
H5: Consumption and sharing of information from social media (Figure 7a, 7c)	–, –	null, –
H6: Active fact-checking and knowledge about how to fact-check (Figures 7b, 6a)	+, +	null, +
H7: Willingness to take COVID-19 precautions (Figure 8a)	+	+
H8: Perceptions of government performance (Figure 8b)	+	+
Comparisons between treatment arms		
Difference between podcast and text	+	–
Difference between empathetic and long podcast	+	+
Difference between long and short podcast	?	null

Note: We merged hypotheses H2 and H3 in our pre-analysis plan into a combined H3; Figure E3a,b in the Supplementary Material report similar results separately. Due to merging these hypotheses and the order that results are presented, H2, H4, H5, H6, and H7 correspond to H5, H6, H4, H7, and H9 in our pre-analysis plan.

engagement. As such, we expected that delivery through an entertaining podcast would be more effective than a text message with the same information, and an empathetic podcast even more so.

MISINFORMATION IN SOUTH AFRICA

Misinformation has been a growing concern in South Africa in recent years, particularly in the context of political and social issues (Posetti, Simon, and Shabbir 2021). In July 2021, for example, national unrest sparked by former president Jacob Zuma's arrest resulted in widespread fake images and posts of destruction and racialized killings appearing on social media, which further exacerbated inter-community tensions, violence, and looting (Allen 2021). During elections, false rumors and conspiracy theories about politicians and political parties have been disseminated to influence voters and to worsen social divisions (International Federation of Journalists 2019). Misinformation has targeted women as frequent subjects, particularly journalists and politicians (Agunwa and Alalade 2022; Wasserman 2020), and has also worsened xenophobic violence in the country (Somdyala 2019).

Since the pandemic's onset in 2020, health misinformation has also increased dramatically. Prominent false claims included COVID-19 not affecting Black Africans, 5G technologies being responsible for the virus, vaccines implanting microchips for government surveillance, and the efficacy of various home remedies and miracle cures (Africa Check 2023). Such pandemic-related misinformation capitalized on citizens' distrust of information provided by their government and perceived political elites (Steenberg et al. 2022). Moreover, misinformation widened health inequality and compliance with government policies; vaccine hesitancy was highest among the most segregated and marginalized communities (Steenberg et al. 2022).

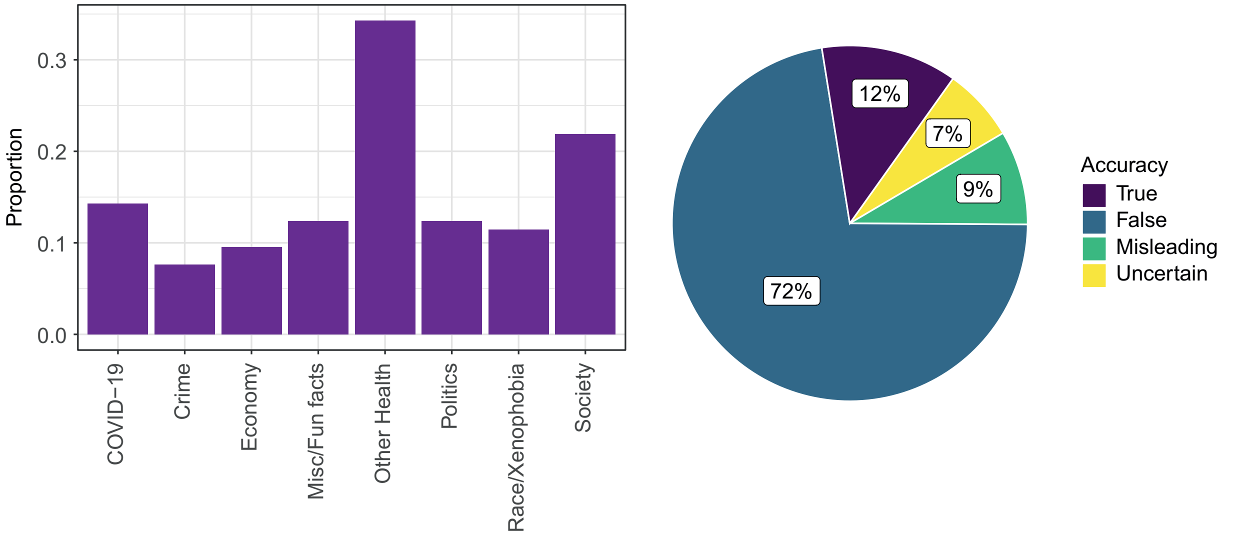
The widespread use of mobile phones and social media platforms like Facebook and WhatsApp in South Africa has fueled the proliferation of misinformation. WhatsApp is a popular choice of communication and news consumption for South African internet users due to its affordability in a country with high data usage costs. In 2021, 88% of South Africans used WhatsApp, and 52% of South Africans used WhatsApp to access news (Newman et al. 2021). However, WhatsApp has also become a breeding ground for misinformation, and its negative impacts worsened during the COVID-19 pandemic (Kazeem 2020).

To combat misinformation, civil society organizations have developed fact-checking tools and initiatives to verify the accuracy of the information circulating on social media. Africa Check is a prominent example: since its founding in 2012, the South African nonprofit has focused its efforts on verifying claims made by public figures and popular content that appears online or on social media. Since 2019, Africa Check has partnered with the podcasting firm Volume to produce a biweekly podcast—entitled "What's Crap on WhatsApp?"—which debunks three locally viral pieces of misinformation each episode in an engaging investigative style. As podcast consumption in South Africa is fast-growing, Africa Check's misinformation podcast seeks to capture a broader audience through an accessible audio format.

RESEARCH DESIGN

To understand the constraints on *consumption* and *internalization* that potentially limit fact-checking's effectiveness, we implemented a six-month field

FIGURE 1. Biweekly Fact-Checked Content



(a) Topics fact-checked

(b) Accuracy of facts checked

Note: Fact-check categories in (a) were coded independently by an undergraduate research assistant. Examples of fact-checks within each category are provided in Section B.1 of the Supplementary Material. Accuracy categories in (b) are provided by Africa Check's fact-checking.

experiment that varied participants' access to different forms of Africa Check's fact-check programming.[3] During the study period, Figure 1 shows that most fact-checks related to (usually false) claims about politics, health issues, and broader social issues. Political fact-checks tended to debunk incendiary claims relating to government corruption or incompetence; health fact-checks often focused on debunking myths and false cures related to COVID-19 and other health conditions (such as diabetes, high blood pressure, or HIV) as well as the adoption of new technologies purportedly harmful to health.

Each fact-check summarized the claim being examined before then explaining Africa Check's process for verifying the claim (which typically involved investigating its source, cross-checking, and consulting experts) and ultimately concluding whether the claim was true or false. Our study evaluates whether sustained exposure to such fact-checks—and how the fact-checks are conveyed—helps individuals to become more discerning of the content they consume in general, changes what information they consume and what they do with it, and ultimately affects citizens' attitudes and behaviors.

Participant Recruitment

Following a brief pilot, the research team recruited participants from social media for the study from across South Africa between October 2020 and September 2021.[4] Facebook advertisements were used to recruit adult Facebook users in 21 "batches" on a rolling basis (typically once every two weeks) for a research study on misinformation in South Africa.[5] Individuals were eligible to participate if they were at least 18 years old, lived in South Africa at the time of recruitment, had a South African phone number, understood English, and used WhatsApp. We restricted our recruitment to social media users due both to their higher anticipated exposure to misinformation as well as the relative feasibility of collecting survey responses (any in-person enumeration would have been challenging due to the COVID-19 pandemic).

Eligible participants then completed a baseline survey administered by the research team via a WhatsApp chatbot (see Figure C1b in the Supplementary Material). The baseline survey recorded participants' demographic characteristics, attitudes regarding misinformation, knowledge about misinformation and current affairs, trust and consumption of different information sources, information verification and sharing behavior, and COVID-19 knowledge and preventative behavior. 11,672 individuals

[3] Data files are available online through the APSR Dataverse (Bowles et al. 2025).

[4] Some participants were therefore recruited—and treated—during the campaign season for the 2021 municipal (local government) elections, which took place on November 1, 2021.

[5] Figure C1a in the Supplementary Material shows an example recruitment ad. Ads were targeted at individuals who did not follow Africa Check's Facebook page, and were stratified at the province-gender-age level to increase representativeness. Few users above 50 years old were targeted, given their lower use of social media. Section A.1 of the Supplementary Material provides additional information on recruitment.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

completed the baseline survey and 8,947 satisfied the conditions necessary to enroll in the study.[6]

This pool of participants was 28 years old on average, and mostly urban (76%), female (61%), and educated (89% report receiving secondary education). Figure C2 in the Supplementary Material compares this sample with nationally representative data from 2018 round of the Afrobarometer survey. While this sample is systematically different from the *overall* broader population, it is similar in terms of observables to the relevant Afrobarometer subgroup who report ever using social media, with only modest differences in age, gender, and education observed.

Treatment Assignment and Delivery

Participants were randomly assigned to either a control group that received no fact-checks or one of four treatment conditions that varied how fact-checks were conveyed. Africa Check delivered these treatments through separate WhatsApp lists created specifically for this intervention, and all treated participants received the same three fact-checks via WhatsApp once every two weeks for six months.[7] Figure 2 provides an example of how the text and three versions of the podcast differed in their presentation of a single fact-check. Table B1 in the Supplementary Material provides additional examples of specific fact-checks, along with examples of wording for the treatment variants.

Fact-Check Treatment Variants

We first varied whether the fact-checks were disseminated through a short text message or a podcast. The *Text* condition simply provided a one-sentence summary of each fact-check, together with a clickable link to an article on Africa Check's website assessing the disputed claim. These messages enabled consumers to quickly learn the veracity of viral online claims without reading the articles, and also to access articles for each of the claims separately.

The three podcast conditions delivered the fact-checks in a more entertaining but longer-form way. In each variant, two narrators explained the veracity of each claim and how they verified the claims in a lively and conversational tone.[8] Among those receiving podcasts, we further varied how costly or empathetic the content was. The default *Long* podcast—which Africa Check disseminates to its regular subscribers—generally lasted 6–8 minutes, while the *Short* podcast cut some discussion of how the claims were verified to reduce the podcast to 4–6 minutes in length. The *Empathetic* podcast augmented the *Long* podcast with empathetic language emphasizing the narrators' understanding of how fear and concern about family and friends might lead individuals to be fooled by misinformation; Section B.3 of the Supplementary Material provides examples of empathetic additions.

Once assigned, treated participants were informed about the mode of dissemination for their fact-checks. 7,331 participants saw their treatment assignment; the residual 1,615, which was balanced across treatment arms, selected out of continued engagement with the study after completing the baseline survey. Treatment was then delivered via Africa Checks' WhatsApp account every fortnight for six months to treated participants, while control participants received no further information from Africa Check.

Incentives to Consume Fact-Checks

To understand organic demand for fact-checks and stimulate engagement among participants lacking interest, we further varied the provision of financial incentives for treated participants to consume Africa Check's fact-checks. Specifically, a randomly selected 83% of treated participants received short monthly quizzes covering recent fact-checks (*fact-check quizzes*). All control participants and the remaining treated participants received quizzes asking about popular culture (*placebo quizzes*). Regardless of quiz type, participants knew in advance that they would receive greater payment for completing these optional monthly quizzes if they answered a majority of quiz questions correctly; see Section A.4 of the Supplementary Material for details. Figure C3 in the Supplementary Material shows that participants who received their treatment regularly took these interim quizzes, with similar rates of quiz participation across treatment arms.

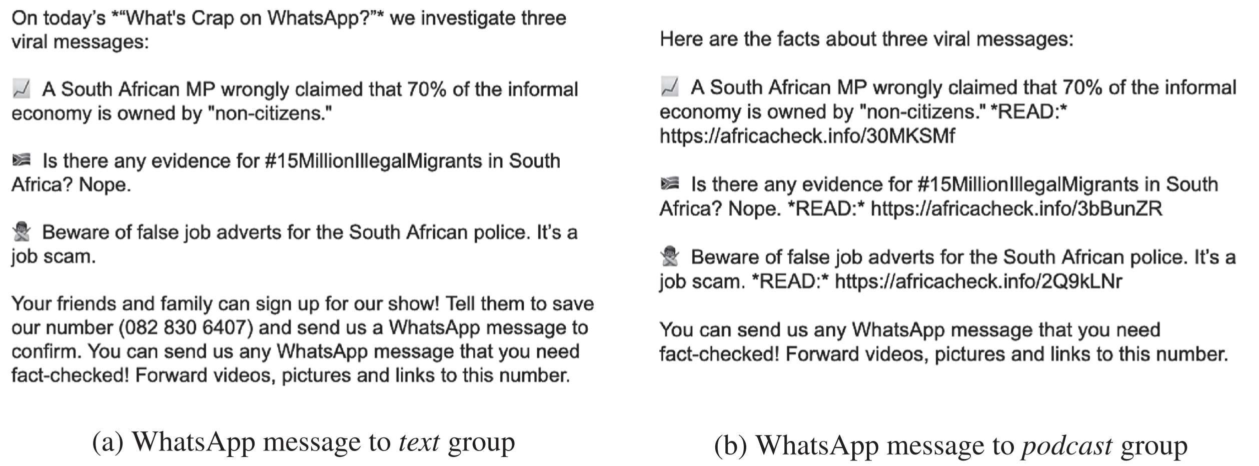
These quizzes were administered by the research team through a different WhatsApp account from the Africa Check account used for treatment delivery. To minimize the risk that the fact-check quizzes would treat participants directly, we did not provide participants with the correct answers or tell them which questions they answered correctly. In line with prior studies adopting similar designs (e.g., Chen and Yang 2019), the quizzes should therefore be construed as generating variation in participants' instrumental incentives to engage with their treatments without constituting an independent source of information in their own right.

[6] Participants were required to send a WhatsApp message to an Africa Check-managed phone number and add that number to their phone contacts to receive a small financial incentive for completing the survey; this was necessary for Africa Check to be able to deliver treatment information to participants through its WhatsApp broadcast lists. Further, we added three simple attention checks (see Section A.1 in the Supplementary Material) to screen out low-quality respondents. Participants had to respond to all attention check questions correctly and not complete the survey in less than eight minutes to enroll in the study.

[7] Although Africa Check delivered these treatments, both the Africa Check team and researchers finalized the wording of the messages to ensure the integrity of treatment variants.

[8] Although participants that received podcasts also received an initial text message similar to the *Text* condition without the links to the articles, their treatment arm was explained as consuming a podcast. Since this instruction was always the most recent, it is likely that participants perceived this intervention as costlier to engage with relative to just reading text information.

FIGURE 2. Example of a Single Fact-Check for Each Treatment Arm



On today's *"What's Crap on WhatsApp?"* we investigate three viral messages:

📈 A South African MP wrongly claimed that 70% of the informal economy is owned by "non-citizens."

🇿🇦 Is there any evidence for #15MillionIllegalMigrants in South Africa? Nope.

👮 Beware of false job adverts for the South African police. It's a job scam.

Your friends and family can sign up for our show! Tell them to save our number (082 830 6407) and send us a WhatsApp message to confirm. You can send us any WhatsApp message that you need fact-checked! Forward videos, pictures and links to this number.

(a) WhatsApp message to *text* group

Here are the facts about three viral messages:

📈 A South African MP wrongly claimed that 70% of the informal economy is owned by "non-citizens." *READ:* https://africacheck.info/30MKSMf

🇿🇦 Is there any evidence for #15MillionIllegalMigrants in South Africa? Nope. *READ:* https://africacheck.info/3bBunZR

👮 Beware of false job adverts for the South African police. It's a job scam. *READ:* https://africacheck.info/2Q9kLNr

You can send us any WhatsApp message that you need fact-checked! Forward videos, pictures and links to this number.

(b) WhatsApp message to *podcast* group

Host 1: Let's jump straight into our first fact-check today. *It was made during the South African state of the nation address on 16 February 2021.* It's the type of claim that can get a lot of people worked up and even angry.

Host 2: Vuyolwethu Zungula, president of the African Transformation Movement, focused on the country's informal economy. This is what he had to say: [AUDIO]

Host 1: *Wow! That's a massive claim.* That 70% of the informal economy is being controlled by foreign nationals. When South Africans – especially those that are unemployed and struggling - hear stats like this it gets their blood boiling.

Host 2: That's right. They can feel like resources are being taken away from them when there isn't much to go around!

Host 1: *How accurate is it?*

Host 2: Zungula provided Africa Check with a number of documents in support of his claim. However, only one included a mention of 70%. This was a report on an interview on talk radio station 702.

Host 1: But the article stated that "foreigners...run about 70% of informal stores in South Africa" — not that they own or control 70% of the informal economy.

Host 2: *The origin of the statistic can be traced to a small survey conducted by Minanawe Marketing saying that 70% of informal retailers were owned by foreigners. These informal stores, known as spaza shops, sell essential household items.*

Host 1: *Zaheera Jinnah, a research associate at Wits University's African Centre for Migration and Society in Johannesburg, told Africa Check that migrant workers are increasing in number but they still constitute a small share of the labour force. What do the stats actually show?*

Host 2: An analysis by the [*short*: Wits University's African Centre for Migration and Society in Johannesburg] [*long:* centre] found that around 20% of people working in the informal economy in 2017 were born outside the country. In Gauteng 19.7% of business owners had migrated to South Africa from another country.

Host 1: *Jinnah warned that "data-light and emotion-heavy" comments like these were likely to "stoke fear and intolerance."*

Host 2: So this claim is incorrect?

Host 1: Yep, it's crap!

(c) Podcast script

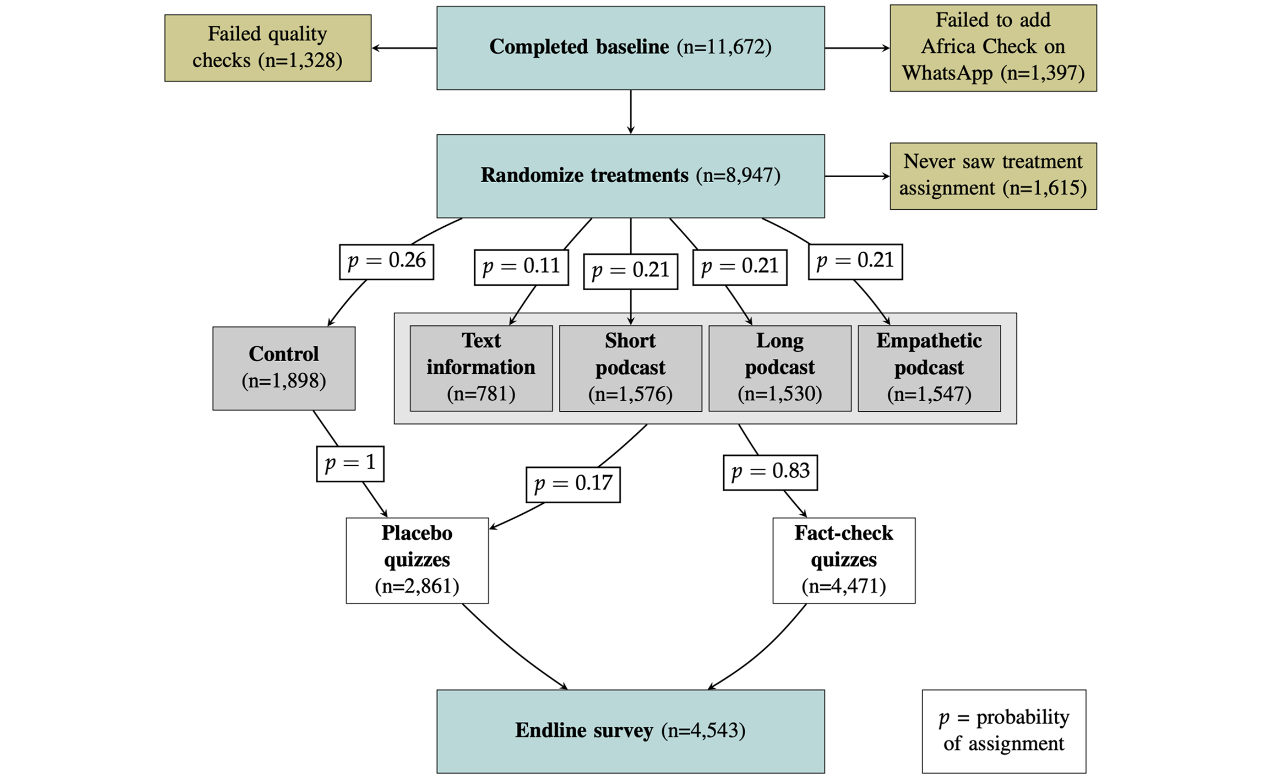
Summary of Interventions

Figure 3 summarizes the overall research design, noting the share of participants assigned to control and each treatment arm as well as the share cross-randomized to fact-check versus placebo quizzes. For each recruitment batch, treatment conditions were randomly assigned within blocks of individuals with similar demographics, social media consumption patterns, trust towards different news sources, and misinformation knowledge.[9] Section A.5 of the Supplementary

[9] We assigned more of the sample to the podcast treatments relative to the text information treatment to improve our statistical power to detect differences across the more similar podcast treatment conditions. In addition to the four main treatment arms, we cross-randomized whether the WhatsApp messages delivering each treatment variant included text

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

FIGURE 3. Overview of Treatment Assignments


Failed quality checks (n=1,328)
Completed baseline (n=11,672)
Failed to add Africa Check on WhatsApp (n=1,397)
Randomize treatments (n=8,947)
Never saw treatment assignment (n=1,615)
$p = 0.26$
$p = 0.11$
$p = 0.21$
$p = 0.21$
$p = 0.21$
Control (n=1,898)
Text information (n=781)
Short podcast (n=1,576)
Long podcast (n=1,530)
Empathetic podcast (n=1,547)
$p = 1$
$p = 0.17$
$p = 0.83$
Placebo quizzes (n=2,861)
Fact-check quizzes (n=4,471)
Endline survey (n=4,543)
p = probability of assignment

Note: The main treatment arms include a pure *Control*, a *Text*-only treatment, a *Short* (4-6 min) podcast, a *Long* (6-8 min) podcast, and an *Empathetic* variant of the long podcast. Participants were additionally incentivized to consume particular content through optional monthly quizzes, relating either to the treatment information (*Fact-check quizzes*) or pop culture (*Placebo quizzes*).

Material provides a discussion of ethical considerations and risks of study participation, which we considered to be minimal.

Outcome Measurement

After six months, each participant completed an endline survey administered by the research team. Those participants who reached the endline ($n = 4,543$) were highly engaged, taking an average of 88% of the monthly quizzes.[10] To uniformly measure fact-check consumption and internalization, we embedded a final quiz relating to Africa Check's recent fact-checks in the endline survey, even if participants had been assigned to placebo quiz incentives during the treatment period. Along with other measures of treatment engagement and internalization, the endline survey measured our primary outcomes: trust in media, attention to content, and discernment of content truth; information consumption, verification, and sharing behaviors; and attitudes and behaviors relating to COVID-19 and politics. Our main analyses aggregate indicators within each of these groups into inverse covariance weighted (ICW) indexes to limit the number of outcomes considered and increase statistical power (Anderson 2008). Table 2 describes each index component, provides summary statistics, and notes the figures in which each outcome is presented. Section A.6 of the Supplementary Material explains how we deal with missing data and justifies some differences from our pre-specified outcome measures.[11]

Estimation

We estimate intent-to-treat (ITT) effects of different combinations of treatment arms relative to our control group. Specifically, we estimate the following pre-specified OLS regressions:

priming the importance of fact-checking for social good. We report the effects of this further encouragement to consume the fact-checks in Section B.4 of the Supplementary Material, where we show that participants assigned to the social prime consumed fact-checks at indistinguishable rates but experienced greater internalization. Given its assignment was orthogonal to the main treatments, our results pool across participants that were and were not primed.

[10] On average, endline respondents received a total of 155 Rand (9.74 USD) through all components of the study.

[11] These differences are quite minimal, but include our decision to exclude questions about WhatsApp itself from indexes relating to social media (given WhatsApp's usage in treatment delivery) and our combination of pre-specified indexes relating to perceived misinformation on and trust in these other social media platforms (due to their high conceptual and empirical overlap).

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

TABLE 2. Outcome Variables

Outcome variable	Variable definitions	Mean	SD	Range
Consumption of fact-checks (H1)				
Podcast take-up (Figure 4a)	How often listen to podcasts (never - all the time)	3.24	1.25	[1,5]
	Included “What’s Crap on WhatsApp” in selection of podcasts listened to	0.41	0.49	{0,1}
Treatment knowledge (Figure 4b)	Number of correct responses from 6 questions on fact-checked content	2.75	1.56	[0,6]
Intended future take-up (Figure 4c)	Stay subscribed (or start subscribing) to “What’s Crap on WhatsApp”	0.83	0.37	{0,1}
	Requested Africa Check’s fact checking content	0.85	0.36	{0,1}
	Requested Africa Check reminders to pay attention to misinformation	0.71	0.45	{0,1}
	Requested vaccine info from Africa Check	0.72	0.45	{0,1}
Discerning fact from fiction (H2–H4)				
Discernment between true and fake news stories (Figure 5a)	Alcohol decreases ability to fight COVID-19 infections (not at all likely - very likely) [true]	3.51	1.27	[1,5]
	Almost 100% of workers in South Africa are foreign (not at all likely - very likely) [false] [–]	2.89	1.31	[1,5]
	COVID-19 spreads by a person’s mouth or nose (not at all likely - very likely) [true]	4.45	0.91	[1,5]
	Matriculation scores to be inflated (not at all likely - very likely) [false] [–]	3.11	1.34	[1,5]
Skepticism of conspiracy theories (Figure 5b)	AIDS intentionally created (not at all likely - very likely)	3.69	1.37	[1,5]
	Nelson Mandela died in 1985 (not at all likely - very likely)	3.82	1.38	[1,5]
	COVID-19 vaccines used to implant chips (not at all likely - very likely)	3.70	1.34	[1,5]
	Vaccines used to reduce world’s population (not at all likely - very likely)	3.72	1.34	[1,5]
Knowledge of verification methods (Figure 6a)	How to avoid being misled: Ask other people [–]	0.13	0.34	{0,1}
	How to avoid being misled: Seek information from reputable organizations	0.36	0.48	{0,1}
	Verification strategies: Ask experts	0.42	0.49	{0,1}
	Verification strategies: Ask themselves [–]	0.88	0.32	{0,1}
	Verification strategies: Check source popularity [–]	0.63	0.48	{0,1}
	Verification strategies: Talk to others [–]	0.82	0.38	{0,1}
	Verification strategies: Use reverse image searches	0.16	0.36	{0,1}
	How to verify info: Ask people I know through WhatsApp [–]	0.82	0.39	{0,1}
	How to verify info: Ask people I know in person [–]	0.71	0.46	{0,1}
	How to verify info: Ask people I don’t know well on WhatsApp group [–]	0.91	0.29	{0,1}
	How to verify info: Ask people I know by posting on social media [–]	0.87	0.33	{0,1}
	How to verify info: Submit a fact-checker request	0.21	0.40	{0,1}
	How to verify info: Go to fact-checker	0.49	0.50	{0,1}
	How to verify info: Use the internet to fact-check yourself	0.46	0.50	{0,1}
Attention to content on social media (Figure 6b)	How often pay close attention to social media information	3.83	1.05	[1,5]
	How to avoid being misled: Pay close attention to source	0.39	0.49	{0,1}
	How important to verify social media information	4.05	1.22	[1,5]
Trust in social media (besides WhatsApp) (Figure 6c)	Likely to be true: Information from other social media (e.g., Facebook, Twitter) (all fake - all truthful)	2.83	0.75	[1,5]
	Trust the most for information: Other social media (e.g., Facebook, Twitter)	0.16	0.37	{0,1}
	Trust: Information from other social media (e.g., Facebook, Twitter) (strongly distrust - strongly trust)	2.88	1.04	[1,5]
Information consumption, verification, and sharing (H5 and H6)				
Social media consumption (Figure 7a)	Regularly go to for news: Other social media (Facebook, Twitter)	0.42	0.49	{0,1}
Active verification (Figure 7b)	How often verify content seen on social media (never - always)	3.83	1.10	[1,5]

(*Continued*)

TABLE 2 (*Continued*)

Outcome variable	Variable definitions	Mean	SD	Range
Sharing (Figure 7c)	How often share social media content shared by others (never - always)	2.83	1.11	[1,5]
Attitudes and behaviors relating to COVID-19 and government (H7 and H8)				
COVID-19 beliefs and preventative behaviors (Figure 8a)	Number of days stayed home in the past week	4.20	2.27	[0,7]
	Number of days visited others indoors in the past week [–]	4.18	2.10	[0,7]
	Number of days wore mask in the past week	5.26	2.36	[0,7]
	View COVID-19 as a fake disease (strongly disagree - strongly agree) [–]	4.36	1.11	[1,5]
	View on COVID-19 lockdown (definitely necessary - definitely unnecessary) [–]	3.21	0.92	[1,4]
	Trust that COVID-19 vaccines in South Africa are safe (strongly distrust - strongly trust)	3.89	1.37	[1,5]
	Would take available COVID-19 vaccine (strongly disagree - strongly agree)	3.49	1.54	[1,5]
Views and attitudes about the government (Figure 8b)	How well national government is performing in general (very badly - very well)	2.38	1.20	[1,5]
	How well national government is handling the COVID-19 pandemic (very badly - very well)	3.09	1.22	[1,5]
	Likely to be true: Information from politicians and government officials (all fake - all truthful)	3.02	0.95	[1,5]
	Trust the most for information: Government officials	0.30	0.46	{0,1}
	Trust the most for information: Politicians and other public figures	0.13	0.34	{0,1}
	Trust: Information from politicians and government officials (strongly distrust - strongly trust)	2.89	1.20	[1,5]
	Vote for regional incumbent in a parliamentary election held tomorrow	0.23	0.42	{0,1}
	Vote for national incumbent in a parliamentary election held tomorrow	0.21	0.41	{0,1}

Note: These descriptive statistics underlie all survey variables used in results and figures presented in Section "Results." The final column represents the full (integer) range of value options in our survey for each question. Variables followed by [–] indicate that variable has been reversed for use in index before providing summary statistics.

$$Y_{ib} = \alpha_b + \beta Y_{ib}^{pre} + \gamma \mathrm{X}_{ib}^{pre} + \boldsymbol{\tau} \mathrm{T}_{ib} + \varepsilon_{ib}, \qquad (1)$$

where Y_{ib} is an outcome for respondent i from block b, T_{ib} is the vector of individual treatment assignments, α_b are randomization block fixed effects, Y_{ib}^{pre} is the baseline analog of the outcome (where feasible), and X_{ib}^{pre} is a vector of predetermined baseline covariates selected separately for each outcome variable via cross-validated LASSO. The vector $\boldsymbol{\tau}$ captures the ITT effect of each treatment condition. Reflecting the individual-level randomization, robust standard errors are used throughout.

We focus on two pre-specified approaches to combining treatment conditions: (i) a pooled specification, where we pool all text and podcast fact-check conditions; and (ii) a disaggregated specification, where we examine the *Text*, *Short* podcast, *Long* podcast, and *Empathetic* podcast conditions separately. The principal deviation from our preregistered specifications is our decision to pool the treated participants that received placebo quiz incentives into a single group (*Placebo incentives*).[12]

For inference, we use one-sided t tests to evaluate hypotheses where we pre-specified a directional hypothesis in Table 2. Otherwise, or in cases where the pre-specified direction is the opposite of the estimated treatment effect, we use two-sided t tests.[13]

We estimate intent-to-treat effects, rather than the local average treatment effects of consuming fact-checks, for several reasons. First, we consider this to be the quantity of theoretical and policy relevance. Our theoretical framework considers potential trade-offs in how fact-checking interventions might shape participants' consumption of corrective information *and* their impacts conditional on consumption. Because we cannot force consumption outside of the lab, understanding the net effect of such interventions—while parsing potential differences in uptake—is then the relevant quantity for policy as well. Second, our treatment conditions could affect relevant outcomes through

[12] We had pre-specified that such individuals would be pooled with groups receiving the *Text*, *Short*, *Long*, or *Empathetic* treatment arm. This ultimately made less sense due to relatively low engagement with fact-checks among participants assigned to placebo quizzes (see Figure 4).

[13] Our index outcomes alleviate some inferential concerns relating to multiple testing. Furthermore, in Table F16 in the Supplementary Material we aggregate our ICW indexes into five meta-indexes, according to the figure in which they appear, and implement a Benjamini–Hochberg correction to account for multiple testing across both treatment coefficients within a specification and outcomes varying between them.

causal pathways that extend beyond just consumption of fact-checks, rendering the exclusion restriction difficult to defend in an instrumental variable analysis. Third, we lack a measure of uptake that does not rely on participants' self-reported consumption of Africa Check's fact-checks.[14]

Finally, we validate the research design in several ways.[15] First, we find no evidence of differential attrition across treatment arms. Table C1 in the Supplementary Material shows balance in the probability of completing the endline survey over treatment conditions.[16] Second, treatment conditions are well balanced across baseline survey covariates in the endline sample.[17] As Table C2 in the Supplementary Material shows, a joint F-test only fails to reject the null hypothesis that the mean of all characteristics are equal to zero at the 10% significance level. Third, we assess the possible concern that demand effects drive our main effects in Section A.7 of the Supplementary Material. As discussed there, we focus on factual outcomes less susceptible to survey response biases, consider such biases to be unlikely to account for differences *between* treatment groups, and find it improbable that biases would affect only the subset of outcome families where we find consistent treatment effects.

RESULTS

We focus on four sets of outcomes. First, we assess how treatment assignment shaped participants' attention to, and consumption of, the fact-checks. Second, we consider whether our sustained intervention improved participants' capacity to discern true and false information *not* covered by the fact-checks. Third, to understand the extent to which individuals reduced their exposure to misinformation, we examine participants' broader media consumption behaviors. Fourth, in line with the fact-checks' topical focus, we evaluate broader impacts on participants' attitudes towards the government and their COVID-19 beliefs and behaviors.

We present results from both the pooled treatment specification and the disaggregated treatment specification. Given our use of ICW indexes to aggregate similar outcome variables, treatment effect estimates reflect standard deviation changes relative to the control group. Our graphical results plot 90% and 95% confidence intervals in each figure; the lower panels provide p-values from tests of differences in effect size between particular treatment arms. Tables F1–F13 in the Supplementary Material report the regression estimates underlying our figures as well as unstandardized estimates for each index component.

Consumption of Fact-Checks

In line with hypothesis H1, we find substantial and sustained levels of fact-check consumption in Figure 4. The upper panel of Figure 4a demonstrates that podcast listenership increased by 0.65 standard deviations across pooled podcast treatment conditions ($p < 0.01$). For our most direct metric of intervention take-up, Table F1 in the Supplementary Material shows that participants assigned to podcasts became 36 percentage points more likely to report listening to the WCW podcast relative to the control group (or participants assigned to text messages) by endline. With respect to text consumption, only around 11% individual webpage links sent as part of the biweekly text messages were clicked by study participants, although the fact-check's conclusion was always conveyed in the WhatsApp message itself.[18]

To capture the extent to which participants paid attention to their assigned treatments, and address the concern that treated respondents over-reported their consumption of the podcast, we consider two behavioral measures of engagement. First, consistent with the debunking aspect of the intervention, Figure 4b demonstrates that the average treated respondent receiving fact-check quiz incentives increased the number of questions about Africa Check's recent fact-checks that they answered correctly on the endline survey by 0.41 standard deviations ($p < 0.01$). This increased the probability of answering such a question correctly from 0.4 to 0.5.

Second, to measure intent to engage with the fact-checks once the modest incentives were removed, we asked participants whether they wished to receive fact-checks, reminders to pay attention to fake news, or COVID-19 vaccine information from Africa Check after the six months of financial incentives concluded. The results in Figure 4c show that treated respondents with incentives to consume fact-checks became 0.21

[14] While we are able to measure the overall frequency with which relevant URL links were clicked, which is relevant for some treatment conditions, we observe this only at the link rather than the individual level.

[15] Because participants are scattered across the country and make up a tiny fraction of the South African population, the stable unit treatment value assumption is likely to hold.

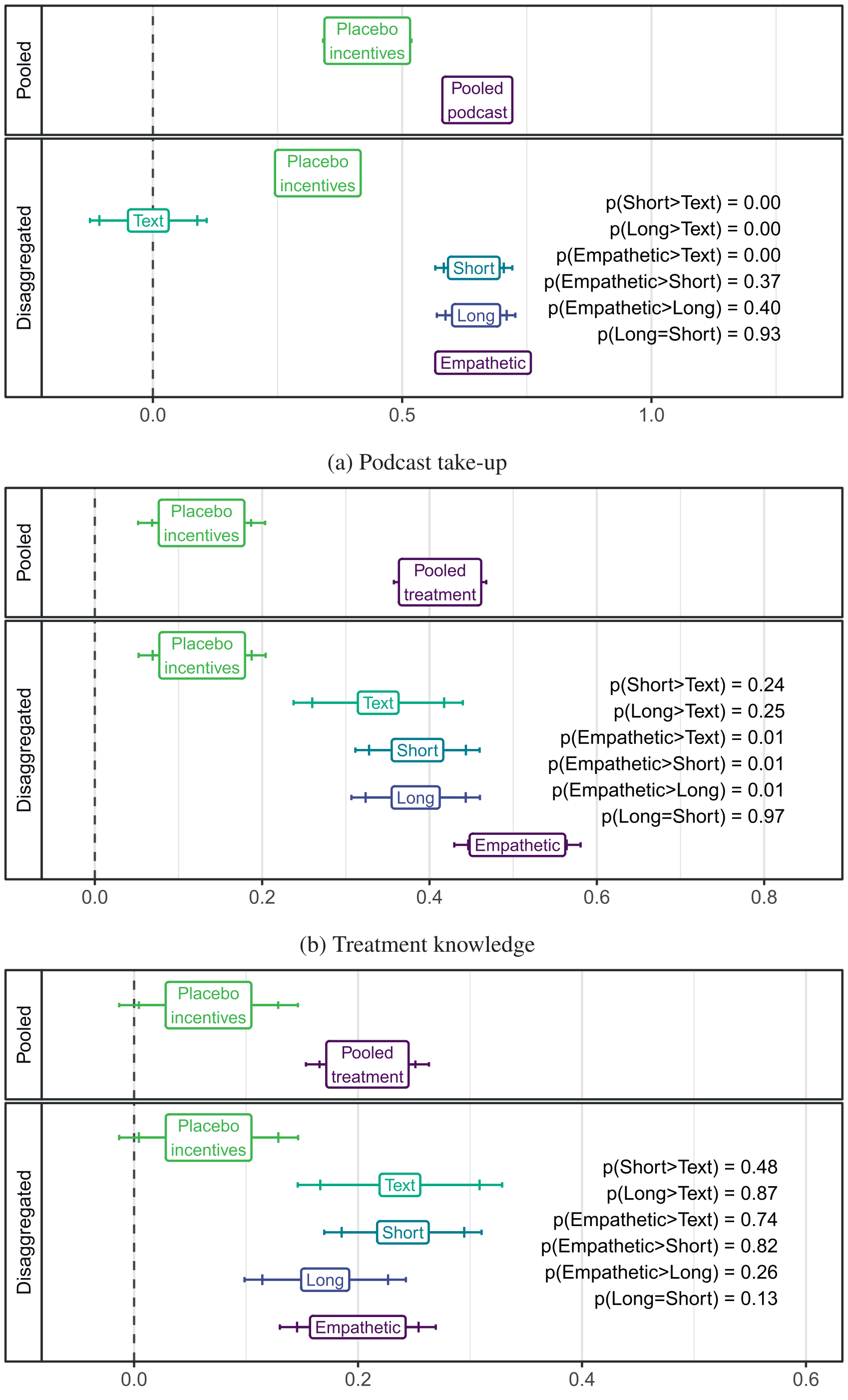
[16] Overall attrition rates from baseline to endline are nearly 50%. These attrition rates owe to the six-month study duration, our use of relatively small financial incentives to induce continued engagement, and our survey enumeration through a WhatsApp chatbot. Participants who dropped out during the study were broadly similar to those who took the endline, aside from being slightly younger and more likely to be male.

[17] Our balance checks include socio-demographic covariates such as gender, age, location, and education level. We also test for balance in covariates that potentially affect susceptibility to misinformation, as identified by the fact-checking literature (see above). These include baseline digital literacy, media consumption and sharing behaviors, beliefs about COVID-19 and misinformation on social media, as well as trust in friends, media outlets, social media, and reputable organizations.

[18] Our experimental design sought to reduce monetary costs associated with data usage by providing participants with a link to a verified source of information, thereby reducing the need for them to search for this information on their own. However, the relatively low click rate suggests that simply clicking a link to a website outside of WhatsApp is indeed still a potential barrier. Thus, relaying the main message within the WhatsApp message itself is an important aspect of the intervention.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

FIGURE 4. Treatment Effects on Take-Up



(a) Podcast take-up

(b) Treatment knowledge

(c) Intended future take-up

Note: All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation 1. Top panel excludes *Text* from *Pooled treatment* since they were not sent podcasts; *p*-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals. Tables F1–F3 in the Supplementary Material report regression results for each index and its components; Tables F14 and F15 in the Supplementary Material further include LASSO-selected covariates.

standard deviations more likely to subscribe to Africa Check's content ($p < 0.01$). Table F3 in the Supplementary Material disaggregates the index to show that the probability of treated respondents signing up to receive the WCW podcast after the intervention increased by 14 percentage points from 75%.

However, indicative of the challenges of generating organic demand for corrective information, the treatments that came with placebo quiz incentives resulted in significantly smaller increases in self-reported engagement, knowledge of fact-checks, and intended future take-up. Our results mirror prior findings suggesting that financial incentives can play a key role in activating latent demand for politically salient information (Chen and Yang 2019). An important challenge for fact-checkers is thus to generate appeal at scale. Our finding that financial incentives generated persisting demand suggests that doing so is possible if initial interest can be ignited. Nevertheless, the limited effects on treatment take-up among participants assigned to placebo incentives leads us to henceforth focus on those treated respondents assigned to fact-check quiz incentives, who engaged far more intensively with their assigned treatments.

The lower panel within each subfigure indicates that treatment take-up was fairly uniform across different treatment conditions where participants were assigned to fact-check quiz incentives. We detect no differences between the long, short, and empathetic podcast conditions in self-reported podcast listening in Figure 4a or between these conditions and the text condition in intended future take-up in Figure 4c. We do find that participants assigned to the empathetic condition were somewhat more accurate in answering questions about recent fact-checks at endline than the other treatment conditions. Rather than differences in engagement, this could reflect empathetic content increasing users' information internalization. Overall, any differences in subsequent effects across treatment variants, conditional on the assignment of fact-check quiz incentives, are thus unlikely to reflect differential take-up and consumption rates.

Discerning Fact from Fiction

Having demonstrated significant engagement with the fact-checks, we next turn to the broader consequences of treatment. We first show that sustained exposure to fact-checks increased treated respondents' ability to discern between true and false content *upon exposure.* In line with established approaches to measuring discernment (see Guay et al. 2023; Pennycook and Rand 2021), we showed respondents two true and two fake news stories relating to COVID-19 and government policy decisions, which were *not* covered by any Africa Check fact-check during the study period. We asked respondents to indicate how likely they believed each to be true on a five-point scale ranging from not at all likely to very likely. We then reverse-coded false questions and produced an ICW index measuring respondents' discernment between true and false. Figure 5a's upper panel shows that any treatment with fact-check quiz incentives increased respondents' discernment between true and false information at endline relative to the control group by 0.06 standard deviations ($p < 0.05$); consistent with their limited consumption of the fact-checks, respondents who received placebo quizzes showed little improvement in misinformation discernment relative to the control group. Figure D1 in the Supplementary Material further shows that improved discernment is driven by respondents' greater distrust of false statements rather than their greater trust of true statements, which suggests that treatment did not simply make people more skeptical of everything they see. As the treatment variant tests in the lower panel illustrate, the pooled treatment effect is driven by the text message and empathetic podcast conditions.

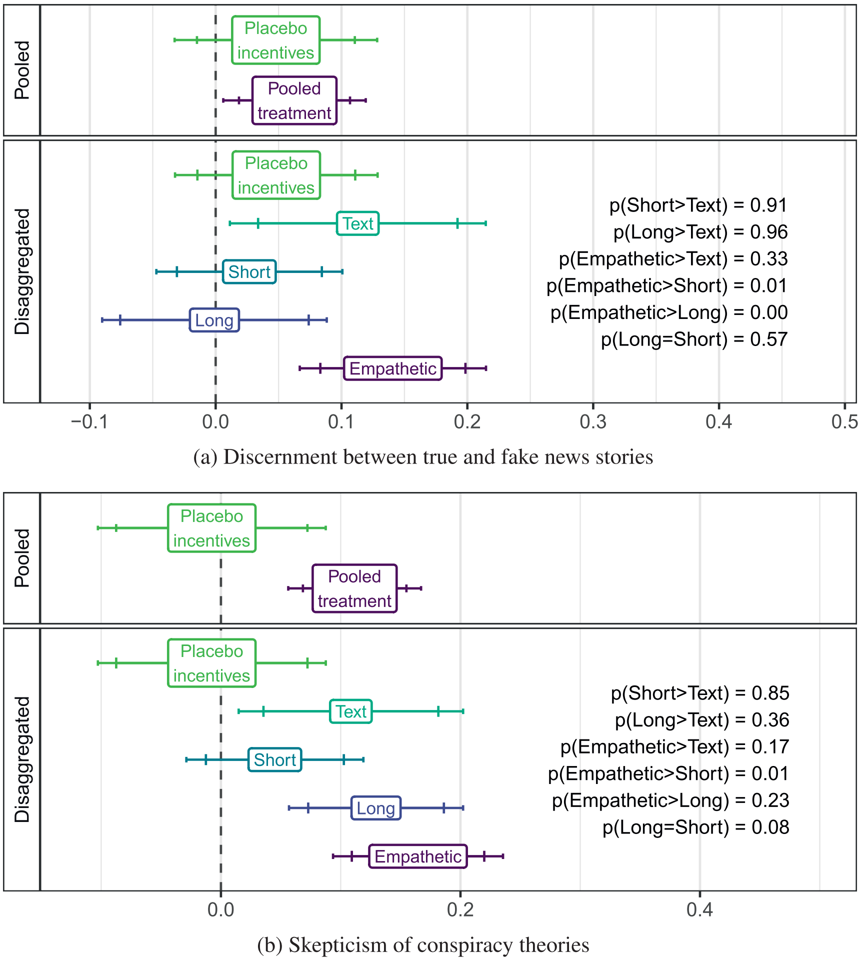
Second, we presented participants with four widespread conspiracy theories *not* investigated by Africa Check (listed in Table 2) and similarly asked respondents to indicate how likely each is to be true. In contrast with the discernment measures above, these outcomes—which were not preregistered because they were less directly related to our treatments—capture skepticism of well-known falsehoods that participants are likely to have encountered in real life (Pennycook and Rand 2020). All responses are reverse-coded such that higher values indicate greater skepticism of the conspiracy theories' likelihood of being true. The upper panel of Figure 5 indicates that pooling treatments with incentives to consume the fact-check quiz increased respondents' skepticism of these conspiracy theories by 0.11 standard deviations, or an average of 0.12 units on our five-point scale ($p < 0.01$). Increased skepticism is driven by the text message and the long and empathetic podcast formats ($p < 0.05$, $p < 0.05$, and $p < 0.01$), which all produced larger effects than the short podcast.

Across participants' ability to discern false from true stories and increased skepticism of conspiracy theories, sustained exposure to fact-checks reduced participants' susceptibility to fake news beyond the fact-checks' narrow content. Supporting hypothesis H4, this suggests that repeated exposure to fact-checks can help to inoculate individuals against misinformation more broadly.

We next consider whether such generalized discernment and skepticism is driven by the broader lessons imparted by Africa Check's fact-checking practices. Suggesting that prebunking is an important component of fact-checks, the upper panel of Figure 6a shows that repeated exposure to fact-checks led respondents to score 0.10 standard deviations higher on our information verification knowledge index ($p < 0.01$), which aggregates 13 items capturing good and bad practices for verifying news. Table F6 in the Supplementary Material disaggregates the index, showing that this effect principally reflects respondents' greater awareness that they can avoid misinformation by relying on reputable sources or consulting fact-checking institutions and cannot effectively verify information simply by asking others. Similar to our discernment outcomes, the lower panel of Figure 6a shows that the text messages and short and empathetic podcast modes of delivery were notably more effective ($p < 0.01$, $p < 0.01$, and $p < 0.05$, respectively) than the standard long podcast.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

FIGURE 5. Treatment Effects on (a) Discernment between Fake and True News and (b) Skepticism of Conspiracy Theories



(a) Discernment between true and fake news stories

(b) Skepticism of conspiracy theories

Note: All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation 1; *p*-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals. Tables F4 and F5 in the Supplementary Material report regression results for each index and its components; Tables F14 and F15 in the Supplementary Material further include LASSO-selected covariates.

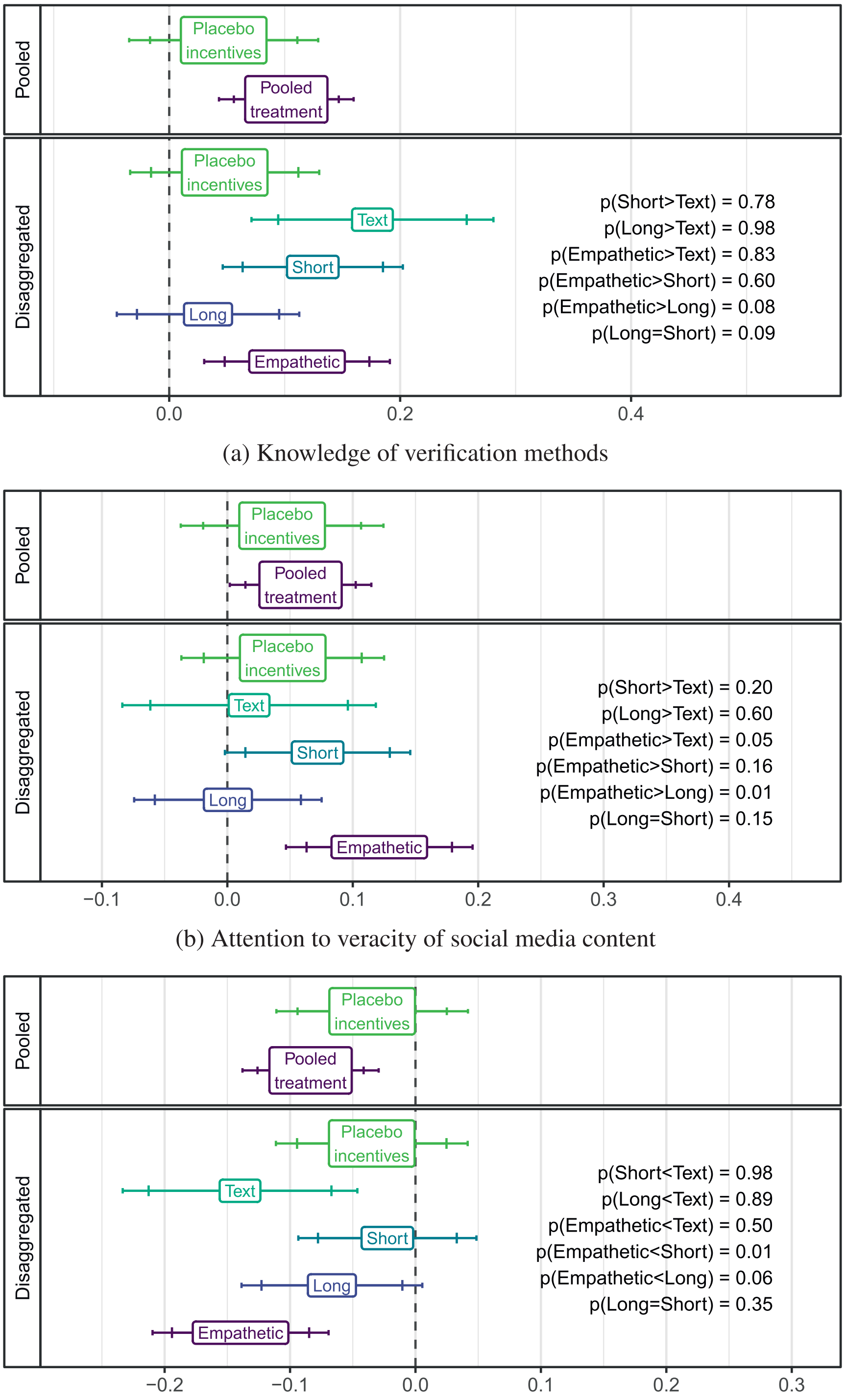
In line with hypothesis H2, these lessons appear to translate into increased attention to the veracity of content on social media. Figure 6b examines an index combining three items: whether respondents listed paying close attention to sources as one of the best ways to avoid being misled by fake news on social media; how important respondents said it is to verify information received on social media; and how frequently they reported paying close attention to content received on social media platforms. The results show a 0.06 standard deviation increase in these measures of perceived importance of paying attention and self-reported attention ($p < 0.05$). This effect is driven, primarily, by respondents becoming 3 percentage points more likely to report that paying close attention to the source of social media content ensures they are not misled by fake news ($p < 0.05$). Comparing across treatment arms, respondents' greater willingness to look twice at content encountered on social media is driven primarily by the empathetic podcast.

Effective inoculation might also reflect greater caution regarding platforms that supply a significant share of misinformation, as predicted by hypothesis H3. Aggregating respondents' assessments of the truthfulness of content on social media platforms with the extent of their trust in such content (other than WhatsApp, through which our fact-checks were delivered), the upper panel of Figure 6c shows that the treatments incentivizing participants to consume fact-checks reduced trust in social media platforms by 0.08 standard deviations ($p < 0.01$).[19] The effect is driven by each component of the index; for example, treatment reduced the share of

[19] We disaggregate this index into participants' perceptions of the truthfulness of social media content, versus their trust in such content, in Appendix Figure E3 in the Supplementary Material and find similar results across each. Figure E4b in the Supplementary Material shows that trust in information from close ties also modestly decreases.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

FIGURE 6. Treatment Effects on News Verification Knowledge, Attention to Veracity of Social Media Content, and Attitudes toward Social Media



(a) Knowledge of verification methods

(b) Attention to veracity of social media content

(c) Trust in social media (besides WhatsApp)

Note: All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation 1; *p*-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals. Tables F6–F8 in the Supplementary Material report regression results for each index and its components; Tables F14 and F15 in the Supplementary Material further include LASSO-selected covariates.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

respondents believing that social media information sources are credible by 15% ($p < 0.01$) and reduced the extent to which content on social media is believed to be true by 0.04 standard deviations ($p < 0.05$). In line with our previous results, the lower panel shows the largest effects for the text message and empathetic podcast delivery formats ($p < 0.01$ and $p < 0.05$, respectively).

Nevertheless, the intervention did not entirely reshape participants' engagement with their information environment. Despite becoming more knowledgeable about verification methods, Figure D2 in the Supplementary Material reports no effect on participants' perceptions about the ease of fact-checking. We also find that beliefs about online sources of information did not carry over to traditional media sources. Figure E4a in the Supplementary Material reports no significant reduction in participants' trust in radio or television news, which tend to be more legitimate sources of information.

Together, these results indicate that sustained access to fact-checks—especially when expressed in a simple text form or conversationally with empathy—increased respondents' capacity to verify suspicious information, generally doubt content on social media *upon exposure*, and ultimately discern misinformation. Further, the heterogeneity across treatment groups, which all received fact-check quiz incentives and experienced similar effects on fact-check consumption, suggests that *how* fact-checked content was conveyed—rather than differential consumption or the quizzes themselves—was responsible for the generally larger effects of the short text messages and more empathetic podcasts.

Information Consumption, Verification, and Sharing

Moving beyond efforts to inoculate participants upon exposure to misinformation, hypothesis H5 assesses whether sustained exposure to fact-checks altered the extent of participants' exposure to and engagement with misinformation in the first place. We first examine treatment effects on a self-reported index of the regularity with which respondents use social media to obtain news (again excluding WhatsApp, through which treatments were delivered). Across our pooled and disaggregated estimations, Figure 7a reports substantively small and consistently statistically insignificant treatment effects ($p = 0.40$). Furthermore, Figure E5 in the Supplementary Material shows that consumption of news from traditional media and close personal ties were also unaffected. Thus, while individuals learned to scrutinize suspect claims and became less trusting of content on social media, the intervention did not shift *where* individuals got their news in the first place. Given that social media platforms are consumed for many purposes beyond acquiring news, this illustrates the supply-side challenge of limiting misinformation exposure.

We similarly observe limited effects on respondents' active efforts to verify the truth of claims encountered outside the study. Failing to reject the null hypothesis for H6, Figure 7a reports no significant increase in how often respondents reported trying to actively verify information they received through social media on a five-point scale from never to always ($p = 0.28$). Figure D3 in the Supplementary Material indicates that, while verification efforts through Africa Check did increase, verification through traditional media was crowded out for all treated participants ($p < 0.01$) and verification via online and social media was crowded out for respondents who were sent fact-checks by text ($p < 0.01$). Along with the increase in verification knowledge observed in Figure 7b, these negligible treatment effects on respondents' verification behavior imply that limited *capacity* to verify news stories might not be the only driver of citizens' limited *efforts* to do so.

While sustained exposure to fact-checks did not affect costly decisions to alter media consumption patterns or actively verify information, greater discernment upon exposure to potential misinformation did translate—for participants that received fact-checks via *Text* or the *Empathetic* podcast—into a reduced propensity to share suspected misinformation. Providing some support for hypothesis H5 with respect to sharing, the lower panel of Figure 7c shows that these participants became around 0.11 standard deviations less likely to report sharing content received via social media ($p < 0.05$), or a 0.1 unit reduction on our five-point scale capturing the frequency with which respondents shared news stories they encounter on social media with others. Thus, in addition to becoming more discerning, sustained treatment may limit viral misinformation outbreaks by making individuals more conscientious about the risks of sharing misinformation once exposed.

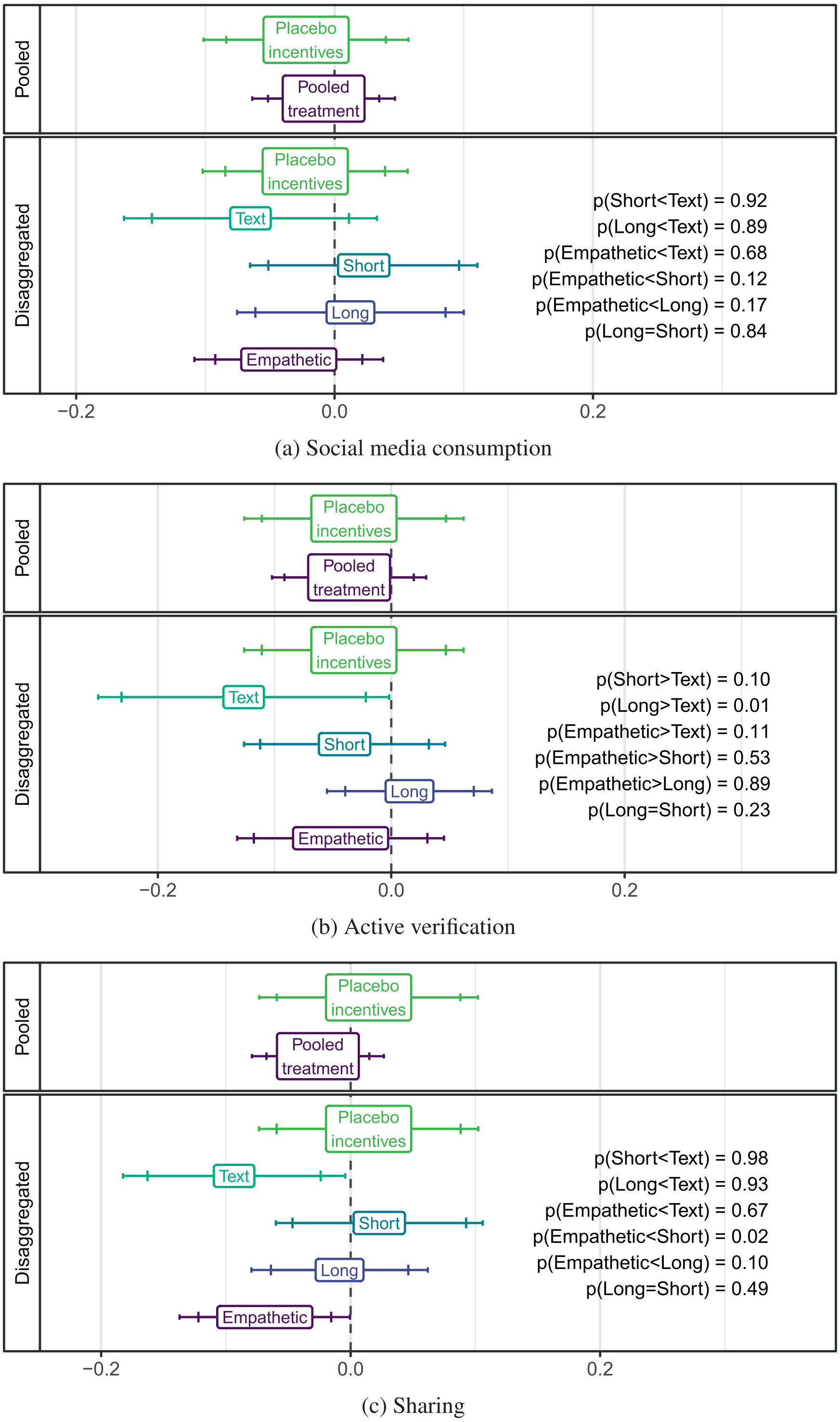
Attitudes and Behaviors Relating to COVID-19 and Government

We finally turn to hypotheses H7 and H8, which assess the political consequences of sustained exposure to fact-checks. A significant share of viral misinformation during the study period related to the COVID-19 pandemic, government officials and policies, and politically salient social issues. By emphasizing false cures or casting doubt on the severity of the pandemic, health-related misinformation risked reducing citizens' compliance with preventative behaviors; exposure to politics-related misinformation would potentially further reduce citizens' trust in formal political institutions. Corresponding fact-checks generally then corrected false claims about COVID-19 and often portrayed incumbent politicians' performance in a more favorable light by casting doubt on outlandish falsehoods.

For our final set of outcomes, we measure effects on indexes of attitudes and self-reported behaviors relating to COVID-19 and politics to assess whether the treatment mitigated the negative downstream consequences typically associated with exposure to misinformation. Since these outcomes are not connected directly to the fact-checks, this enables us to test whether sustained efforts to combat salient misinformation influenced participants' perspectives on public health and politics more broadly.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

FIGURE 7. Treatment Effects on Social Media Consumption, Verification, and Sharing



(a) Social media consumption

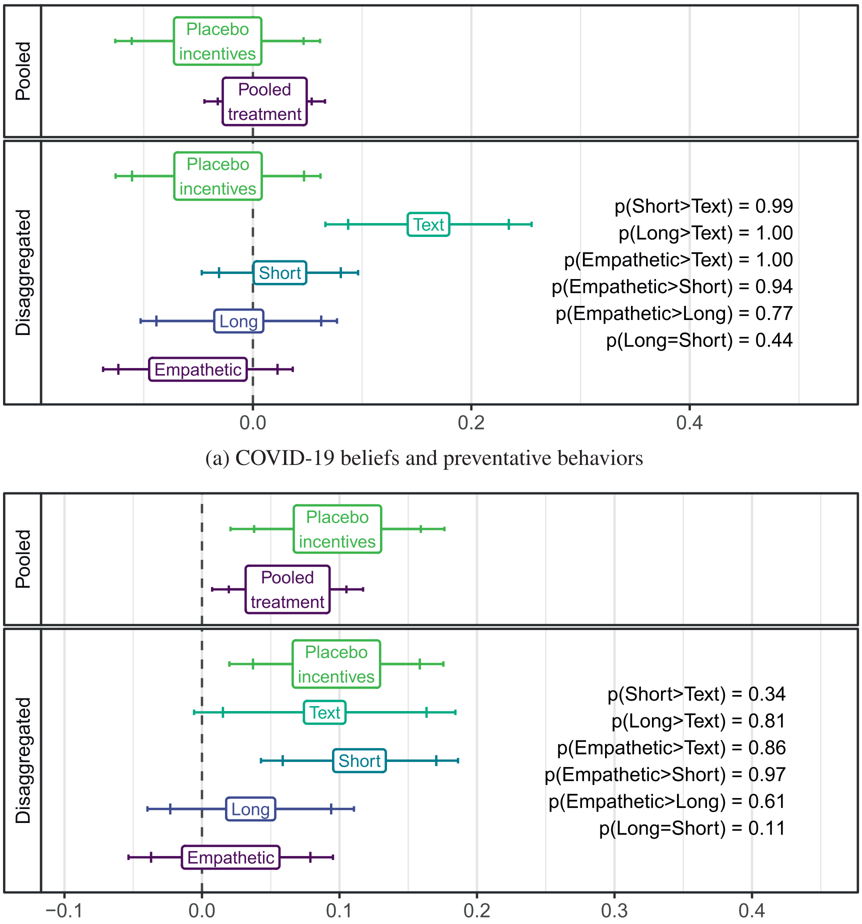
(b) Active verification

(c) Sharing

Note: All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation 1; *p*-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals. Tables F9–F11 in the Supplementary Material report regression results for each index and its components; Tables F14 and F15 in the Supplementary Material further include LASSO-selected covariates.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

FIGURE 8. Treatment Effects on COVID-19 Beliefs and Preventative Behaviors and Views and Attitudes about the Government



(a) COVID-19 beliefs and preventative behaviors

(b) Views and attitudes about the government

Note: All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation 1; *p*-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals. Tables F12 and F13 in the Supplementary Material report regression results for each index and its components; Tables F14 and F15 in the Supplementary Material further include LASSO-selected covariates.

Overall, we detect modest but some significant effects after six months of exposure to fact-checks on such beliefs and behaviors. Figure 8a generally reports no treatment effect on COVID-19 beliefs and preventative behavior for the three podcast treatments with fact-check quiz incentives. However, providing some support for hypothesis H7, we find that fact-checks delivered by short and simple text messages increased an index of health-conscious outcomes associated with COVID-19 by 0.16 standard deviations ($p < 0.01$). Examining the components of the index separately, Table F12 in the Supplementary Material indicates that the effects of the text-only treatment are driven by significant increases in respondents' willingness to comply with government policies by getting vaccinated, wearing a mask, and reducing indoor activity.

In line with hypothesis H8, Figure 8b reports an increase in favorable views toward the government—measured in terms of government performance appraisals, trust in government, and intentions to vote for their region's incumbent party—across treatment conditions. The pooled treatment effect of 0.06 standard deviations ($p < 0.05$) is largely driven by the text message format ($p < 0.1$) and the short podcast ($p < 0.01$). Table F13 in the Supplementary Material shows that these effects are primarily driven by significant increases in the extent to which respondents reported trusting information from politicians and the government most, as well as more favorable evaluations of government performance and willingness to vote for incumbent parties.

These results indicate that broader politically-relevant beliefs and behaviors are harder to move than the capacity to discern fact from fiction. Nevertheless, our findings suggest that the greater discernment and verification knowledge inspired by sustained exposure to fact-

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

checks may start to push individuals to make fact-based judgments in their private and political lives as well. In particular, text messages that can be consumed at little cost appear to help combat misinformation-induced perspectives on polarizing issues.

CONCLUSION

Due to its potentially negative consequences for political and health-related behaviors, misinformation on social media is a growing concern around the globe. Recent studies have advanced our understanding of how to mitigate the consumption of, and susceptibility to, misinformation online. But most struggle to explain how sustained changes in beliefs and behaviors can be achieved beyond single-shot treatments with short-term effects.

In addition to estimating effects of sustained exposure to fact-checks, we explored two key challenges in a world where many factors compete for citizens' attention: how to generate *persistent consumption* of corrective information and how to induce *internalization* of the lessons imparted by fact-check content. Our sustained intervention allowed us to examine whether fact-checking can play both debunking and prebunking roles by correcting existing misinformation and warning participants about future misinformation.[20] Partnering with an existing fact-checking organization, Africa Check, also highlights the relatively low cost and scalability of the intervention.

Our study yields several key conclusions. First, it is feasible to stimulate citizens to consume fact-checking content delivered through WhatsApp. Modest financial incentives helped to induce consumption in our South African sample; once the incentives were removed, treated participants expressed their desire to continue receiving Africa Check's content. Consequently, while organic consumption was difficult to generate at the beginning, an initial push towards consumption may subsequently activate latent demand. Once high-quality fact-checking content is available, policymakers and researchers should therefore focus on generating initial exposure by: (i) finding ways to cultivate citizen demand for fact-checking content, whether by increasing their perceived importance or by capturing spillovers by enhancing their entertainment value; or (ii) using tools like media campaigns, trusted community leaders or online influencers, or school curricula to, at least initially, disseminate fact-checking content widely.

Second, sustained exposure to fact-checks helped to inoculate participants upon exposure to misinformation. While treated participants did not report altering behaviors that limit exposure to misinformation in the first place or active verification efforts, the intervention increased participants' attention to veracity as well as capacity to discern fact from fiction and willingness to act on this by not sharing unverified online content. Since the effects we observe are relatively small in magnitude, it is imperative to increase the efficacy of inoculation efforts beyond the effects we document in this study. Such efforts should complement supply-side efforts targeting the production and promotion of misinformation.

Third, not all treatment arms performed equally: the simple text-only treatment and empathetic podcast treatments were the most effective delivery mechanisms for internalizing fact-checking messages. Our results thus suggest that 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 prioritizes empathizing with consumers.

Finally, our results suggest that combating misinformation can be politically consequential. Although not all types of fact-checks generated significant effects, we find that sustained exposure to fact-checks made citizens somewhat more compliant with government policies and more trusting in incumbent governments. As such, text-based fact-checks that could be consumed almost costlessly helped to reverse two important concerns of the social media age, reduced state capacity and declining faith in government.

While our findings illustrate the potential for sustained fact-checking interventions to make citizens more discerning of the content they consume, several limitations point to avenues for future research. First, although fact-checks were consumed in a natural environment, we still recruited participants for an academic study. Since willing participants are likely to be at least somewhat interested in subscribing to online content, further research should examine whether our findings are larger or smaller in broader populations with lower barriers to opt-in and lower baseline levels of interest in and knowledge of fact-checking. Second, our results suggest that *sustained* exposure to fact checks can more durably reduce participants' beliefs in misinformation and increase their ability to sort fact from fiction. However, our experiment was not designed to identify the length of exposure needed to achieve these effects —an important avenue for future work. Third, our treatment effects were strongest for participants with financial incentives to consume fact-checking content. As such, we demonstrate that the intervention can be effective, but more organically generating repeated exposure to fact-checks remains a critical outstanding research question. Fourth, this study largely relies on self-reported survey responses. Although our findings are unlikely to be driven by experimental demand effects (for reasons discussed above), studies conducted in more naturalistic environments would benefit from linking fact-checking interventions to behavioral outcomes.

In addition, several extensions to the study could improve our understanding of fact-checking and misinformation reduction overall. For example, there is

[20] In Figure D4 in the Supplementary Material, we explored whether treatment effects are heterogeneous across populations with potentially differing levels of digital literacy, as proxied by pre-treatment knowledge of how to verify information and education levels. Overall, we find little consistent evidence of heterogeneous treatment effects, perhaps because our social media-selected sample is overall quite homogeneous. For example, nearly 80% of the sample have secondary but not tertiary education.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

more to be learned about which aspects of fact-checks make citizens more discerning. Our findings point to trust in content's source and a general understanding of when and how to exercise skepticism as key mechanisms, but future research could further break apart the components of a typical fact-check to identify which components to emphasize. In addition, while we find positive effects from sustained exposure to fact-checking, our study design did not allow us to investigate potential variation based on ongoing political events, such as major elections. To understand how increased misinformation and polarization may shift fact checks' efficacy, more research is needed. Finally, while Africa Check is a reputable fact-checking organization, fact-checkers may themselves be poor sources of information in some contexts. Future research may consider the question of how to ensure the credibility of fact-checkers.

South Africa's salient political issues during this period—pandemic responses, identity politics, and economic concerns—are reflected in politics globally and contribute to misinformation worldwide. Our findings show that similar mechanisms that have helped to combat misinformation relating to polarizing issues in the Global North can generalize to Global South contexts, and thus advance existing scholarship on limiting misinformation's potential for adverse effects (Blair et al. 2024; Cook, Lewandowsky, and Ecker 2017; Nyhan 2020; Walter et al. 2020). However, insofar as the problems of misinformation are exacerbated by the use of closed platforms such as WhatsApp, and by lower digital literacy in part due to costs associated with data usage, our study's context highlights that the challenges of misinformation—and need for low-cost and scalable solutions—is especially pressing in the Global South. However, our findings are more optimistic than prior media literacy studies in the Global South (Badrinathan 2021; Guess et al. 2020), suggesting that sustained observational learning via social media platforms themselves can help to combat misinformation.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0003055424001394.

DATA AVAILABILITY STATEMENT

Research documentation and data that support the findings of this study are openly available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/UNOOY0.

ACKNOWLEDGMENTS

We thank Africa Check and Volume, and in particular Kate Wilkinson, Taryn Khourie, and Paul McNally, for their cooperation with this study. We are grateful for helpful comments and feedback from Leticia Bode and Molly Offer-Westort and from participants of MPSA 2022, DIMIS Workshop, BID Workshop at TSE, PSPB Seminar at LSE, GSPP Seminar at UC Berkeley, PED Lunch at Stanford, Trust & Safety Research Conference 2022, and APSA 2022. This study was pre-registered in the Social Science Registry (www.socialscienceregistry.org/trials/7615), which also houses our pre-analysis plan. With thanks to Mert Akan and Rachel Raps for research assistance. The research team is grateful for funding from by the Harvard Data Science Initiative Trust in Science program, the Harvard Center for African Studies, and Facebook Health Partnerships programs. Larreguy gratefully acknowledges funding from the French Agence Nationale de la Recherche under the Investissement d'Avenir program ANR-17-EURE-0010.

FUNDING STATEMENT

This research was funded by the Harvard Data Science Initiative Trust in Science program, the Harvard Center for African Studies, and Facebook Health Partnerships Program.

CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors declare the human subjects research in this article was reviewed and approved by Institutional Review Boards at Columbia (IRB-AAAT2554), Duke (2024-003), Harvard (IRB20-0602), and UC Berkeley (2020-07-13490). The authors affirm that this article adheres to the principles concerning research with human participants laid out in APSA's Principles and Guidance on Human Subject Research (2020).

REFERENCES

Africa Check. 2023. "Fact-Checks." https://africacheck.org/fact-checks.

Agunwa, Nkemakonam, and Temiloluwa Alalade. 2022. "Dangers of Gendered Disinformation in African Elections." WITNESS. https://blog.witness.org/2022/08/dangers-of-gendered-disinformation-in-african-elections/.

Ali, Ayesha, and Ihsan Ayyub Qazi. 2023. "Countering Misinformation on Social Media through Educational Interventions: Evidence from a Randomized Experiment in Pakistan." *Journal of Development Economics* 163: 103108.

Allen, Karen. 2021. "Social Media, Riots and Consequences." *Institute for Security Studies*. https://issafrica.org/iss-today/social-media-riots-and-consequences.

Alt, James E., John Marshall, and David D. Lassen. 2016. "Credible Sources and Sophisticated Voters: When Does New Information Induce Economic Voting?." *Journal of Politics* 78 (2): 327–42.

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484):1481–95.

Argote, Pablo, Elena Barham, Sarah Zuckerman Daly, Julian E. Gerez, John Marshall, and Oscar Pocasangre. 2021. "Messages That Increase COVID-19 Vaccine Acceptance: Evidence from Online Experiments in Six Latin American Countries." *PLOS One* 16 (10): e0259059.

Badrinathan, Sumitra. 2021. "Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India." *American Political Science Review* 115 (4): 1325–41.

Bandura, Albert. 2001. "Social Cognitive Theory of Mass Communication." *Media Psychology* 3 (3): 265–99.

Banerjee, Abhijit, Eliana La Ferrara, and Victor H. Orozco-Olvera. 2019. "The Entertaining Way to Behavioral Change: Fighting HIV with MTV." *National Bureau of Economic Research*. Working Paper Series No. 26096.

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya. 2020. "Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics." *Journal of Public Economics* 182: 104–23.

Baum, Matthew A. 2002. "Sex, Lies, and War: How Soft News Brings Foreign Policy to the Inattentive Public." *American Political Science Review* 96 (1): 91–109.

Baum, Matthew A., and Angela S. Jamison. 2006. "The Oprah Effect: How Soft News Helps Inattentive Citizens Vote Consistently." *Journal of Politics* 68 (4): 946–59.

Berlinski, Nicolas, Margaret Doyle, Andrew M. Guess, Gabrielle Levy, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, et al. 2023. "The Effects of Unsubstantiated Claims of Voter Fraud on Confidence in Elections." *Journal of Experimental Political Science* 10 (1): 34–49.

Blair, Robert A., Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J. Stainfield. 2024. "Interventions to Counter Misinformation: Lessons from the Global North and Applications to the Global South." *Current Opinion in Psychology* 55: 101732.

Bode, Leticia, Emily Vraga, and Melissa Tully. 2020. "Do the Right Thing: Tone May Not Affect Correction of Misinformation on Social Media." *HKS Misinformation Review*. 1 (4). https://doi.org/10.37016/mr-2020-026.

Bowles, Jeremy, Horacio Larreguy, and Shelley Liu. 2020. "Countering Misinformation via WhatsApp: Preliminary Evidence from the Covid-19 Pandemic in Zimbabwe." *PLOS One* 15 (10): e0240005.

Bowles, Jeremy, Kevin Croke, Horacio Larreguy, Shelley Liu, and John Marshall. 2025. "Replication Data for: Sustaining Exposure to Fact-Checks: Misinformation Discernment, Media Consumption, and Its Political Implications." Harvard Dataverse. Dataset. https://doi.org/10.7910/DVN/UNOOY0.

Carey, John M., Andrew M. Guess, Peter J. Loewen, Eric Merkley, Brendan Nyhan, Joseph B. Phillips, and Jason Reifler. 2022. "The Ephemeral Effects of Fact-Checks on COVID-19 Misperceptions in the United States, Great Britain, and Canada." *Nature Human Behaviour* 6 (2): 236–43.

Chan, Man-pui Sally, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. "Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation." *Psychological Science* 28 (11): 1531–46.

Chen, Yuyu, and David Y. Yang. 2019. "The Impact of Media Censorship: 1984 or Brave New World?" *American Economic Review* 109 (6): 2294–332.

Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, et al. 2020. "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media." *Political Behavior* 42: 1073–95.

Cook, John. 2013. "Inoculation Theory." In *The Sage Handbook of Persuasion: Developments in Theory and Practice*, eds. James Price Dillard and Lijiang Shen, 220–36. Thousand Oaks, CA: SAGE.

Cook, John, Stephan Lewandowsky, and Ullrich K.H. Ecker. 2017. "Neutralizing Misinformation through Inoculation: Exposing Misleading Argumentation Techniques Reduces Their Influence." *PLOS One* 12 (5): e0175799.

Flynn, D. J., Brendan Nyhan, and Jason Reifler. 2017. "The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics." *Advances in Political Psychology* 38 (1): 127–50.

Gentzkow, Matthew, Michael B. Wong, and Allen T. Zhang. Forthcoming. "Ideological Bias and Trust in Information Sources." *American Economic Journal: Microeconomics*.

Gesser-Edelsburg, Anat, Alon Diamant, Rana Hijazi, and Gustavo S. Mesch. 2018. "Correcting Misinformation by Health Organizations during Measles Outbreaks: A Controlled Experiment." *PLOS One* 13 (12): e0209505.

Gottlieb, Jessica, Claire L. Adida, and Richard Moussa. 2022. "Reducing Misinformation in a Polarized Context: Experimental Evidence from Côte d'Ivoire." *OSF Preprints*. https://osf.io/6x4wy.

Guay, Brian, Adam J. Berinsky, Gordon Pennycook, and David Rand. 2023. "How to Think about Whether Misinformation Interventions Work." *Nature Human Behaviour* 7 (8): 1231–33.

Guess, Andrew M., Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. "A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India." *Proceedings of the National Academy of Sciences* 117(27):15536–45.

Guess, Andrew M., Nyhan Brendan, and Jason Reifler. 2020. "Exposure to Untrustworthy Websites in the 2016 US Election." *Nature Human Behaviour* 4 (5): 472–80.

Hameleers, Michael. 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 & Society* 25 (1): 110–26.

Henry, Emeric, Ekaterina Zhuravskaya, and Sergei Guriev. 2022. "Checking and Sharing Alt-Facts." *American Economic Journal: Economic Policy* 14 (3): 55–86.

Hopkins, Daniel J., John Sides, and Jack Citrin. 2019. "The Muted Consequences of Correct Information about Immigration." *Journal of Politics* 81 (1): 315–20.

International Federation of Journalists. 2019. "South Africa: Disinformation Is the Biggest Threat to Any Election Process." https://www.ifj.org/media-centre/news/detail/category/africa/article/south-africa-disinformation-is-the-biggest-threat-to-any-election-process

Iyengar, Ananya, Poorvi Gupta, and Nidhi Priya. 2023. "Inoculation Against Conspiracy Theories: A Consumer Side Approach to India's Fake News Problem." *Applied Cognitive Psychology* 37 (2): 290–303.

Jerit, Jennifer, and Yangzi Zhao. 2020. "Political Misinformation." *Annual Review of Political Science* 23: 77–94.

Kalla, Joshua L, and David E. Broockman. 2020. "Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments." *American Political Science Review* 114 (2): 410–25.

Kazeem, Yomi. 2020. "With over 250,000 Cases, Misinformation Is Compromising Africa's Covid-19 Response." *Quartz Africa*, June 22. https://qz.com/africa/1871683/whatsapp-is-a-key-source-of-covid-19-information-for-africans.

Kim, Eunji. 2023. "Entertaining Beliefs in Economic Mobility." *American Journal of Political Science* 67 (1): 39–54.

Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. 2000. "Misinformation and the Currency of Democratic Citizenship." *Journal of Politics* 62 (3): 790–816.

La Ferrara, Eliana. 2016. "Mass Media and Social Change: Can We Use Television to Fight Poverty?" *Journal of the European Economic Association* 14 (4): 791–827.

Lewandowsky, Stephan, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest* 13 (3): 106–31.

Maertens, Rakoen, Jon Roozenbeek, Melisa Basol, and Sander van der Linden. 2021. "Long-Term Effectiveness of Inoculation Against Misinformation: Three Longitudinal Experiments." *Journal of Experimental Psychology: Applied* 27 (1): 1–16.

Marshall, John. 2023. "Tuning in, Voting Out: News Consumption Cycles, Homicides, and Electoral Accountability in Mexico." Working Paper.

Martel, Cameron, Gordon Pennycook, and David G. Rand. 2020. "Reliance on Emotion Promotes Belief in Fake News." *Cognitive*

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press

Research: Principles and Implications 5 (47). https://doi.org/10.1186/s41235-020-00252-3

Martel, Cameron, Mohsen Mosleh, and David G. Rand. 2021. "You're Definitely Wrong, Maybe: Correction Style Has Minimal Effect on Corrections of Misinformation Online." *Media and Communication* 9 (1): 120–33.

Newman, Nic, Richard Fletcher, Anne Schulz, Simge Andı, Craig T. Robertson, and Rasmus Kleis Nielsen. 2021. "Reuters Institute Digital News Report 2021." *Reuters Institute*. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf.

Nyhan, Brendan. 2020. "Facts and Myths about Misperceptions." *Journal of Economic Perspectives* 34 (3): 220–36.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J. Wood. 2020. "Taking Fact-Checks Literally but Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability." *Political Behavior* 42 (3): 939–60.

Nyhan, Brendan, and Jason Reifler. 2015. "Displacing Misinformation about Events: An Experimental Test of Causal Corrections." *Journal of Experimental Political Science* 2 (1): 81–93.

Offer-Westort, Molly, Leah R. Rosenzweig, and Susan Athey. 2024. "Battling the Coronavirus 'Infodemic' Among Social Media Users in Kenya and Nigeria." *Nature Human Behaviour* 8 (5): 823–34.

Pennycook, Gordon, and David G. Rand. 2020. "Who Falls for Fake News? The Roles of Bullshit Receptivity, Overclaiming, Familiarity, and Analytic Thinking." *Journal of Personality* 88 (2): 185–200.

Pennycook, Gordon, and David G. Rand. 2021. "The Psychology of Fake News." *Trends in Cognitive Sciences* 25 (5): 388–402.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. "Shifting Attention to Accuracy Can Reduce Misinformation Online." *Nature* 592: 590–5.

Pereira, Frederico Batista, Natália S Bueno, Felipe Nunes, and Nara Pavão. 2024. "Inoculation Reduces Misinformation: Experimental Evidence from Multidimensional Interventions in Brazil." *Journal of Experimental Political Science* 11 (3): 239–50.

Peterson, Erik, and Shanto Iyengar. 2021. "Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading?" *American Journal of Political Science* 65 (1): 133–47.

Porter, Ethan, and Thomas 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): e2104235118.

Posetti, Julie, Felix Simon, and Nabeelah Shabbir. 2021. "Reporting Elections on the Frontline of the Disinformation War." *Reuters Institute*. https://reutersinstitute.politics.ox.ac.uk/news/reporting-elections-frontline-disinformation-war

Prior, Markus. 2007. *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections*. New York: Cambridge University Press.

Roozenbeek, Jon, and Sander Van der Linden. 2019. "Fake News Game Confers Psychological Resistance Against Online Misinformation." *Palgrave Communications* 5 (1): 1–10.

Servick, Kelly. 2015. "Fighting Scientific Misinformation: a South African Perspective." *Science*, February 15. https://www.science.org/content/article/fighting-scientific-misinformation-south-african-perspective.

Shehata, Adam, David Nicolas Hopmann, Lars Nord, and Jonas Höijer. 2015. "Television Channel Content Profiles and Differential Knowledge Growth: A Test of the Inadvertent Learning Hypothesis Using Panel Data." *Political Communication* 32 (3): 377–95.

Somdyala, Kamva. 2019. "Fake News about Xenophobia on Social Media Aimed at Ruining Brand SA." *News24*, April 03. https://www.news24.com/news24/fake-news-about-xenophobia-on-social-media-aimed-at-ruining-brand-sa-govt-20190403.

Steenberg, Bent, Nellie Myburgh, Andile Sokani, Nonhlanhla Ngwenya, Portia Mutevedzi, and Shabir A. Madhi. 2022. "COVID-19 Vaccination Rollout: Aspects of Acceptability in South Africa." *Vaccines* 10 (9): 1379.

Stroud, Natalie Jomini, Joshua M Scacco, and Yujin Kim. 2022. "Passive Learning and Incidental Exposure to News." *Journal of Communication* 72 (4): 451–60.

Taber, Charles S., and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50 (3): 755–69.

Tewksbury, David, Andrew J. Weaver, and Brett D. Maddex. 2001. "Accidentally Informed: Incidental News Exposure on the World Wide Web." *Journalism & Mass Communication Quarterly* 78(3): 533–54.

Tucker, Joshua A., Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, et al. 2018. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." Report. Hewlett Foundation.

Tully, Melissa, Emily K. Vraga, and Leticia Bode. 2020. "Designing and Testing News Literacy Messages for Social Media." *Mass Communication and Society* 23 (1): 22–46.

Walter, Nathan, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. "Fact-Checking: A Meta-Analysis of What Works and for Whom." *Political Communication* 37 (3): 350–75.

Wasserman, Herman. 2020. "Fake News from Africa: Panics, Politics and Paradigms." *Journalism* 21 (1): 3–16.

Williamson, Scott, Claire L. Adida, Adeline Lo, Melina R. Platas, Lauren Prather, and Seth H. Werfel. 2021. "Family Matters: How Immigrant Histories Can Promote Inclusion." *American Political Science Review* 115 (2): 686–93.

Wood, Thomas, and Ethan Porter. 2019. "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence." *Political Behavior* 41 (1): 135–63.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Zukin, Cliff, and Robin Snyder. 1984. "Passive Learning: When the Media Environment Is the Message." *Public Opinion Quarterly* 48(3): 629–38.

https://doi.org/10.1017/S0003055424001394 Published online by Cambridge University Press