Sustaining Exposure to Fact-checks: Misinformation Discernment, Media Consumption, and its Political Implications[*]

Jeremy Bowles,[†] Kevin Croke,[‡] Horacio Larreguy,[§] Shelley Liu,[¶] John Marshall[‖]

October 2023

Exposure to misinformation can affect citizens' beliefs, political preferences, and compliance with government policies. However, little is known about how to reduce susceptibility to misinformation in a sustained manner outside controlled environments, 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 impactful. However, we find limited effects on news consumption choices or verification behavior. Our results demonstrate that inducing sustained exposure to fact-checks might inoculate citizens against future misinformation, but highlight the difficulty of inducing broader behavioral changes relating to media usage.

[*]With thanks to Africa Check and Volume, and in particular Kate Wilkinson, Taryn Khourie, and Paul McNally, for their cooperation. 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. IRB review granted by Columbia (IRB-AAAT2554), Harvard (IRB20-0602), and UC Berkeley (2020-07-13490). 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, 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.

[†]Department of Political Science and School of Public Policy, University College London.

[‡]Harvard T.H. Chan School of Public Health, Harvard University.

[§]Departments of Economics and Political Science, ITAM.

[¶]Sanford School of Public Policy, Duke University.

[‖]Department of Political Science, Columbia University.

1 Introduction

Misinformation about politics, social issues, and public health is a growing and ubiquitous concern. Such content—defined by its potential to generate misperceptions about the true state of the world—encourages beliefs and behaviors 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 Tironi et al. 2021; Berlinski et al. 2021). 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. forthcoming), misinformation has already been linked to lynchings and mass electoral mobilization in India and racial violence in South Africa (Allen 2021; Badrinathan 2021).

Efforts to limit the potential impact of misinformation typically engage in *debunking* or *prebunking*. 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, derived from 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. forthcoming; 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.

Fact-checking—one popular method of combating misinformation—may present complementarities between debunking and prebunking. Fact-checking most obviously debunks by informing citizens about particular false (and true) claims. However, it should also prebunk by increasing general awareness of misinformation, explaining the logic behind common forms of misinformation, and explaining 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.

However, despite these potential benefits, it is difficult to induce citizens to 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 forced consumption settings (e.g. Porter and Wood 2021), outside of the lab it 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 problem of internalization—even conditional on information consumption (Zaller 1992)—and calls into question fact-checking's efficacy at combating misinformation beyond the lab or online surveys. Moreover, little is yet known about how fact-checking shapes political dispositions beyond those narrowly 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. Twice a week for six months, treated participants in our large rolling sample of social media users were sent three fact-checks via WhatsApp messages. These fact-checks dissected largely false stories that were trending on social media in South Africa in the preceding weeks pertaining to politics, health, and other high-profile topics. 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]

[1]The quizzes did not provide the correct answers, and thus simply provided incentives to consume fact-checks rather than constituting an additional information source. Moreover, since incentives were constant across conditions,

We further examine if, and how, citizens can be induced to engage and internalize fact-checks by randomly varying how the fact-checks were disseminated to participants. These four treatment conditions varied the appeal and cost of consuming the fact-checks, and 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 link to additional information assessing a 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. Third, recognizing limits on time and attention span, we tested an abbreviated 4-6 minutes 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 empathy plays in driving the internalization of information (Gesser-Edelsburg et al. 2018; Gottlieb, Adida and Moussa 2022; Kalla and Broockman 2020).

Our corresponding panel survey establishes 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, even absent the provision of incentives, suggesting that the intervention activated latent demand—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.

we can isolate the effect of different conditions on information internalization upon consumption.

Second, sustained exposure to fact-checks helps to inoculate citizens against misinformation *upon exposure*. Receiving any incentivized form of treatment persistently increased respondents' ability to discern true from false stories relating to politics and public health issues and increased their skepticism towards prominent conspiracy theories—none of which were covered during the intervention. Our results suggest that this may be driven by treated participants' increased understanding of what credible content looks like, their reduced trust in social media, and their greater capacity to verify content for themselves. Nevertheless, the treatments did not impact the amount of news that participants consumed from social and traditional media—and thus their risk of being exposed to misinformation—and their verification behavior. These results suggest that sustained exposure to fact-checks primarily combats misinformation by increasing skepticism upon exposure to such content, rather than by altering the type of content individuals consume in the first place.

Third, comparisons across treatment variants 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 can be effective particularly when it includes emotive appeals to increase the resonance of corrective information with consumers.

Our study adds to the comparatively limited, albeit growing, body of work studying interventions to hinder misinformation in Global South contexts (cf. Ali and Qazi 2023; Badrinathan 2021; Pereira et al. forthcoming; Porter and Wood 2021; Gottlieb, Adida and Moussa 2022). In particular, a recent comprehensive review of misinformation studies noted that more than 80% focused on Global North countries, which "highlights the challenges of drawing conclusions about effective strategies for countering misinformation in the Global South" (Blair et al. 2023, 3). Our study's findings thus help to validate the benefits of inoculation even in settings where consumers have variable media literacy levels or where they face high data costs of independently validating information they find on social media platforms.

The unusually sustained aspect of our field experimental intervention, along with the richness of our research design, mean that our findings advance our 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 that single-shot media literacy interventions can effectively prebunk misinformation (Maertens et al. 2021; Pereira et al. forthcoming; Roozenbeek and Van der Linden 2019; cf. Badrinathan 2021; Hameleers 2022). We also contribute to this literature by showing interventions conducted outside controlled research environments can be effective when citizens are motivated to consume fact-checks. By further measuring a particularly broad array of outcomes, we establish that the enduring effects of our prebunking 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 hope for demand-side interventions, this finding simultaneously emphasizes the need for complementary supply-side change.

Second, our findings illuminate the theoretical mechanisms required for fact-checks to be impactful at scale. In line with inventive studies seeking to "gamify" digital literacy lessons (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 emphatically—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 competitive 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 translate into such outcomes, misinformed beliefs could instead reflect partisan cheerleading with more limited political impact (Jerit and Zhao 2020). By demonstrating that WhatsApp-based text messages regularly conveying fact-checks both increase faith in the incumbent government and reported compliance with its policies, we show that (combating) misinformation can have durable political consequences. We are not aware of any studies that have previously established this connection outside of lab settings. Our results thus corroborate the perception that modern polities should be concerned about misinformation's potentially corrosive effects on state capacity and political accountability.

2 When might fact-checking be effective?

Within developing country settings, 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 2022). 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. forthcoming). 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 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 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—which is derived from inoculation theory—posits that people can be "inoculated" against misinformation in general when they are consistently warned about misinformation's existence and are equipped with tools to identify it (Cook 2013; Martel, Pennycook and Rand 2020). Common prebunking interventions include warning labels or digital literacy training (e.g. Badrinathan 2021; Cook, Lewandowsky and Ecker 2017; Offer-Westort, Rosenzweig and Athey 2022; Pereira et al. forthcoming; Tully, Vraga and Bode 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 highlight the broader threat of misinformation, explain how misinformation can be debunked using reliable sources and fact-checking techniques, and simultaneously explain the faulty logic behind certain false claims. Ultimately, fact-checking may not only debunk the misinformation it discusses but also prebunk new misinformation by increasing consumers' media literacy, thereby generating awareness about how to spot misinformation and engage in fact-checking themselves.

Fact-checks can potentially then combat misinformation in two main ways. First, misinformation's impact could be reduced *upon exposure* as people become more discerning of, and also more equipped to verify, what they are consuming. Even if their overall exposure to misinformation is not affected, internalization of the lessons from fact-checks may nevertheless ensure that individuals become more skeptical of the misinformation—and, ideally, more trusting of truthful information—they encounter on social media or elsewhere. Second, they could *reduce exposure* to misinformation by teaching individuals how to recognize—and thus avoid—potential sources of such misinformation. Because fact-checks also educate people about which types of sources are legitimate information providers, they may start consuming more reputable sources.

Although a number of studies experimentally demonstrate fact-checking's promise (see Nyhan 2020), these studies also have important limitations (Flynn, Nyhan and Reifler 2017; Walter et al. 2020). First, existing work primarily relies on one-shot interventions and often forces participants' exposure to fact-checks in lab or survey environments. Outside these settings, however, citizens who allocate their time across a wide array of activities often choose not 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 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 (Taber and Lodge 2006; Peterson and Iyengar 2021). Furthermore, existing research has tended to find only short-term success in combating the *specific* pieces of misinformation that the 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 has negative implications for fact-checking's potential benefits for media literacy.

2.1 Improving the efficacy of fact-checks

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

2.1.1 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 emotive content on social media, misinformation-correcting interventions that are quicker to digest for users might induce more consumption than interventions that take longer to parse and understand. Given that internalization depends on initial consumption, easier-to-consume fact-checks may ultimately prove to be more effective in increasing audience reach and awareness.

Another potential solution is to make fact-checking content more *engaging*. Research on "edutainment" demonstrates how delivering information in more interesting and entertaining ways positively affects consumption, information recall, beliefs, and behaviors (e.g. Baum 2002; Baum and Jamison 2006; Kim 2023; La Ferrara 2016). Notably, Banerjee, La Ferrara and Orozco-Olvera (2019) found that exposure to television programming helped to increase awareness of HIV and health behaviors in Nigeria. Furthermore, Roozenbeek and Van der Linden (2019) and Maertens et al. (2021) 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.

2.1.2 Enhancing internalization

Sustained exposure may mitigate some of the shortcomings associated with the internalization of fact-checking interventions. First, by increasing the volume of content consumed, sustained exposure might reduce the likelihood that fact-checking content is crowded out by other content.[2] Second, internalization of media literacy lessons may require longer and more frequent exposure (Guess et al. 2020; Tully, Vraga and Bode 2020). While individual fact-checks may teach viewers about certain warning signs, consistent fact-checking content can help to build up an arsenal of

[2] In addition, when consumers receive fact-checks consistently, they are more likely to be aware of the prevalence of misinformation, leading them to become more careful about what they read.

reliable strategies and misinformation logics, which encourage critical thinking skills and equip people to be more discerning media consumers. Third, sustained exposure could enhance users' trust in the fact-checking source (Gentzkow, Wong and Zhang 2021), which may in turn increase internalization (Alt, Marshall and Lassen 2016).

The mode by which fact checks are delivered might also 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. Finally, 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 out-group exclusion, while Williamson et al. (2021) finds 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, the trade-off between the optimal level of detail and tone of delivery may instead *decrease* fact-checks' effectiveness.

2.2 Theoretical expectations

Together, we anticipate that sustained exposure to fact-checking ought to combine aspects of both debunking and prebunking for misinformation correction. By enhancing consumers' consumption and internalization of corrective information, their ability to discern true from false information online and knowledge of verification techniques should increase, while reducing the extent of their trust in, and consumption of, social media content. Citizens exposed to such fact-checking interventions over a sustained period could then either learn to identify and discern misinformation, and also verify it, upon exposure, or otherwise change their behaviors which affect exposure to misinformation in the first place. To the extent that misinformation typically focuses on salient false claims about politics or public policy, sustained exposure to fact-checks might then induce improved perceptions of government performance and compliance with its policies.

Understanding *how* to effectively increase organic consumption and internalization, however, is 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. Appendix A.5 discusses our pre-specified expectations relating to this trade-off for our study, including that interventions leveraging "edutainment" or more empathetic content would be most effective by enhancing internalization at the potential cost of initial consumption.

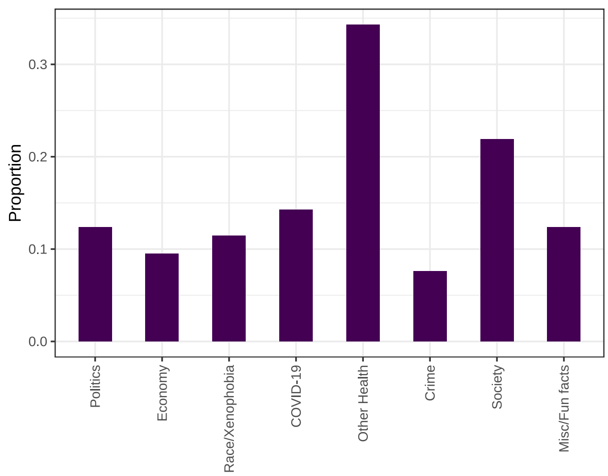
3 Misinformation in South Africa

Misinformation has been a growing concern in South Africa in recent years, particularly in the context of political and social issues (Reuters Institute 2021). In July 2021, for example, national unrest sparked by former president Jacob Zuma's arrest resulted in widespread faked 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 2021). Misinformation has targeted women, particularly journalists and politicians (Agunwa and Alalade 2022; Wasserman 2020), and has also worsened xenophobic violence in the country (News24 2019).

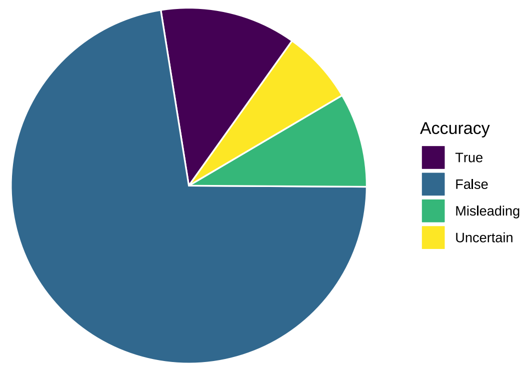
Since the pandemic's onset in 2020, health misinformation has also increased dramatically. From rumors that COVID-19 did not affect Black Africans, to vaccines implanting microchips for government surveillance, to home remedies and miracle cures (Africa Check 2023), pandemic-related misinformation capitalized on deep citizen distrust of information provided by their government and perceived global elites (Steenberg et al. 2022). Such misinformation has 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 stands out as 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 have only worsened during the COVID-19 pandemic (Quartz Africa 2020).

To combat rising quantities of 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 also 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 entertaining investigative style.



(a) Topics fact-checked



(b) Accuracy of facts checked

Figure 1: Biweekly fact-checked content

Notes: Fact check categories in figure (a) are coded independently by an undergraduate research assistant. Accuracy categories in figure (b) are provided by Africa Check's fact-checking.

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.

4 Research design

To understand the constraints on *consumption* and *internalization* that potentially limit fact-checking's effectiveness, we implemented a six-month field experiment that varied participants' access to different forms of Africa Check's fact-checking programming. During the study period, Figure 1 shows that most of these fact-checks related to (generally false) claims about politics, health issues, and broader social issues. Political fact-checks tended to debunk incendiary claims relating to government corruption or incompetence, while health fact-checks often focused on debunking myths and false cures related to COVID-19. Appendix B.1 provides specific examples.

4.1 Participant recruitment

Following a brief pilot, we recruited participants from social media for the study from across South Africa between October 2020 and September 2021, with participants recruited in 21 "batches" on a rolling basis (typically once every two weeks). Facebook advertisements were used to recruit adult Facebook users for a research study on misinformation in South Africa (see Appendix Figure C1a).[3] 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 extremely challenging due to the COVID-19 pandemic).

Eligible participants then completed a baseline survey administered via a WhatsApp chatbot (see Appendix Figure C1b). The baseline survey recorded participants' demographic characteristics, attitudes regarding misinformation, baseline 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 completed the baseline survey and 8,947 satisfied the conditions necessary to enroll in the study.[4]

This pool of participants was 28 years old on average, and mostly urban (76%), female (61%), and educated (89% report receiving secondary education). Appendix Figure C2 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.

[3]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. See Appendix A.1 for additional information on recruitment.

[4]Participants were further 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 simple attention checks (see Appendix A.1) to screen out low-quality respondents.

4.2 Treatment assignment and delivery

Our sample of participants were randomly assigned to either a control group that received no fact-checks or one of four treatment conditions. All treated participants received the same three fact-checks via WhatsApp once every two weeks for six months; Appendix B.1 provides examples of specific fact-checks. However, the fact-checks were delivered in different ways across treatment conditions.

4.2.1 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.[5] 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; Appendix B.2 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 1,616, which was bal-

[5] 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.

anced 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 two weeks for six months to treated participants, while control participants received no further information from Africa Check.

4.2.2 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 Appendix A.3 for details. Participants who received their treatment regularly took these interim quizzes, with similar rates of quiz participation across treatment arms (see Appendix Figure C3).

The fact-check quizzes did not provide participants with the correct answers or tell them which questions they answered correctly. Further, these quizzes were administered through a different WhatsApp account from the Africa Check account used for treatment delivery. 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.

4.2.3 Summary of interventions

Figure 2 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

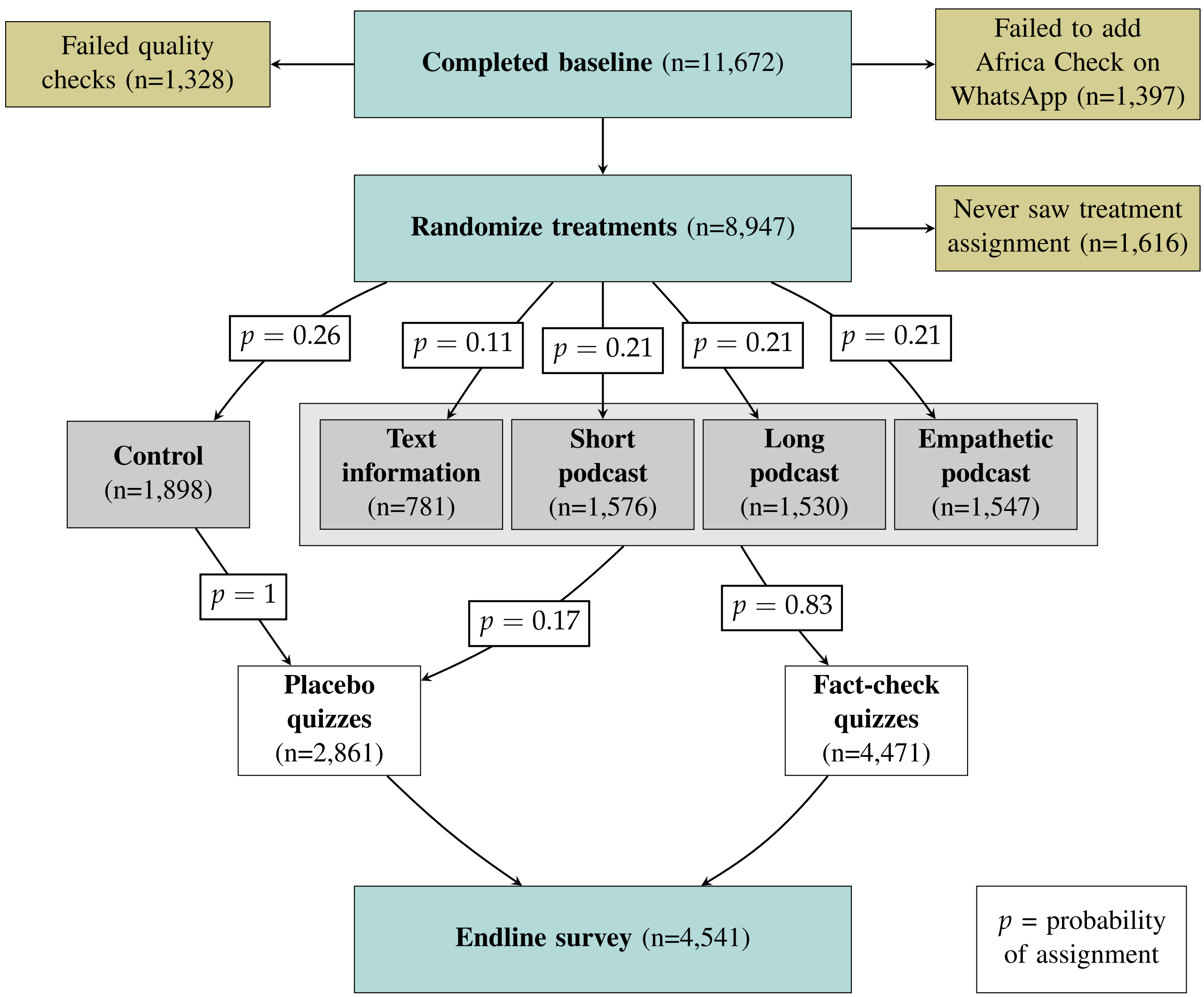


Figure 2: **Overview of treatment assignments**
The main treatment arms include a pure *Control*, a *Text*-only treatment, a *Short* (4-6min) podcast, a *Long* (6-8min) podcast, and an *Empathetic* variant of the long podcast. Participants were additionally incentivized to consume particular content through optional monthly quizzes, relating either to the treatment information (*Fact-check quizzes*) or pop culture (*Placebo quizzes*).

of individuals with similar demographics, social media consumption patterns, trust towards different news sources, and misinformation knowledge.[6] Section A.4 provides a discussion of ethical considerations relating to the interventions and risks of study participation, which we considered to be minimal.

4.3 Outcome measurement

After six months, each participant completed an endline survey. Those participants reaching the endline ($n = 4,541$) were highly engaged, taking an average of 88% of the monthly quizzes.[7] To uniformly measure fact-check consumption and internalization, we embedded a final quiz which related to Africa Check's recent fact-checks 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: discernment of content truth, verification knowledge, and trust in media; information consumption, verification, and sharing patterns; and attitudes and self-reported 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). Appendix Table C1 provides definitions and summary statistics for each index component, while Appendix A.6 notes how we deal with missing data and justifies some differences from our pre-specified outcome measures.

[6] 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 priming the importance of fact-checking for social good. We report the effects of this further encouragement to consume the fact-checks in Appendix B.3, 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.

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

4.4 Estimation

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

$$Y_{ib} = \alpha_b + \beta Y_{ib}^{pre} + \boldsymbol{\gamma} \mathbf{X}_{ib}^{pre} + \boldsymbol{\tau} \mathbf{T}_{ib} + \varepsilon_{ib}, \tag{1}$$

where Y_{ib} is an outcome for respondent i from block b, $\mathbf{T}_{ib}$ 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 $\mathbf{X}_{ib}^{pre}$ 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.

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 *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*).[8] Reflecting the individual-level randomization, robust standard errors are used throughout. For inference, we use one-sided t tests to evaluate hypotheses where we pre-specified a directional hypothesis (see Appendix A.5). Otherwise, or in cases where the pre-specified direction is the opposite of the estimated treatment effect, we use two-sided t tests.

Our primary treatment effect estimates then comprise the effect of assignment to fact-check quizzes as well as assignment to a given treatment condition for those participants assigned to fact-check quizzes. As noted above, since the quizzes did not provide the correct answers, these primarily increased participants' incentives to consume their assigned podcast or text messaging rather than constituting an independent informational intervention. Because such incentives should

[8] 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 3).

be constant across different conditions assigned to fact-check quizzes, comparisons *across* treatment arms isolates differences owing only to variation in the modes of fact-check delivery.

We focus on intent-to-treat effects, rather than the local average treatment effect arising from an instrumental variable (IV) estimation, 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. Exactly 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 potentially have multiple causal pathways that affect relevant outcomes, rendering the exclusion restriction difficult to defend in an IV design. Additionally, we lack a cleanly defined measure of uptake that does not rely on participants' self-reporting.[9]

We validate the research design in several ways.[10] First, we examine differences in the probability of completing the endline survey by treatment arm. Appendix Table C2 shows balance in attrition across treatment conditions.[11] Second, we conduct balance tests across baseline survey covariates in the endline sample. As Appendix Table C3 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 Appendix A.7. 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.

[9]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.

[10]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.

[11]Overall attrition rates from baseline to endline are nearly 50% but are indistinguishable across treatments. 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 during the study were broadly similar to those who took the endline, aside from being slightly younger and more likely to be male.

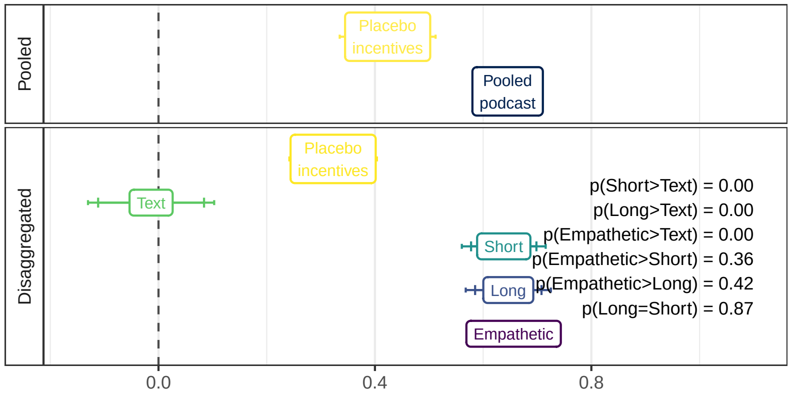
5 Results

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

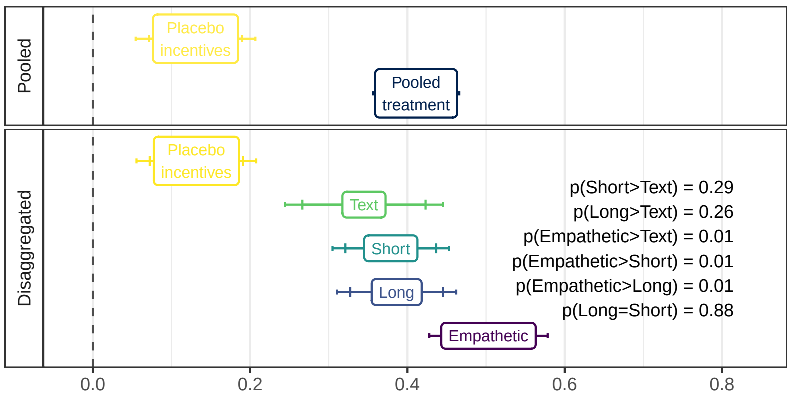
We present results from both the pooled treatment specification and the disaggregated treatment specification. Given our use of index variables, treatment effect estimates reflect standard deviation changes relative to the control group. Our graphical results plot 90% and 95% confidence intervals in each figure; the lower panels provide p-values from tests of differences in the effects between particular treatment arms, which test for our directional hypotheses noted above. Appendix Tables F1-F13 report the regression estimates underlying our figures as well as unstandardized estimates for each index component.

5.1 Consumption of fact-checks

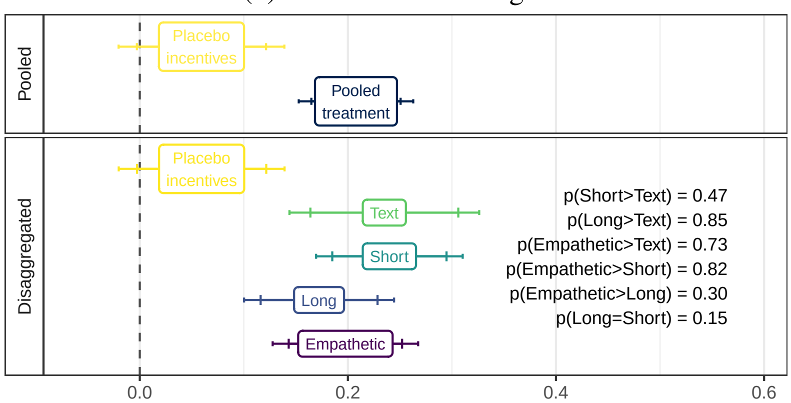
We find substantial and sustained levels of fact-check consumption in Figure 3. The upper panel of Figure 3a demonstrates that podcast listenership increased by 0.65 standard deviations ($p < 0.01$) across pooled podcast treatment conditions ($p < 0.01$). For our most direct metric of intervention take-up, Appendix Table F1 shows that podcast-assigned participants became 36 percentage points more likely to report listening to the WCW podcast relative to the control group (or text-assigned participants) by endline. Self-reported treatment take-up is balanced across podcast conditions. With respect to text consumption, only around 11% of the total number of individual webpage links sent out were clicked by study participants, although the fact-check's conclusion was always conveyed in the WhatsApp message itself.



(a) Podcast take-up



(b) Treatment knowledge



(c) Intended future take-up

Figure 3: Treatment effects on take-up

Notes: All outcomes are standardized inverse covariance-weighted indexes (see Table C1): (a): how often reports listening to podcasts and reports listening to WCW; (b) number of fact-check quiz questions answered correctly out of 6; (c) indicators for wanting future Africa Check (AC) vaccine info, AC fact-checks, AC reminders, and to subscribe to WCW. 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 of Figure 3a 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. Regression tables including all index components provided in Appendix Tables F1-F3. Regression table of ICW indexes including all LASSO-selected controls provided in Appendix Tables F14-F15.

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 3b 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 continue receiving information from Africa Check after the six months of financial incentives concluded. The results in Figure 3c show that treated respondents with incentives to consume fact-checks became 0.2 standard deviations more likely to subscribe to Africa Check's content ($p < 0.01$). Appendix Table F3 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%.

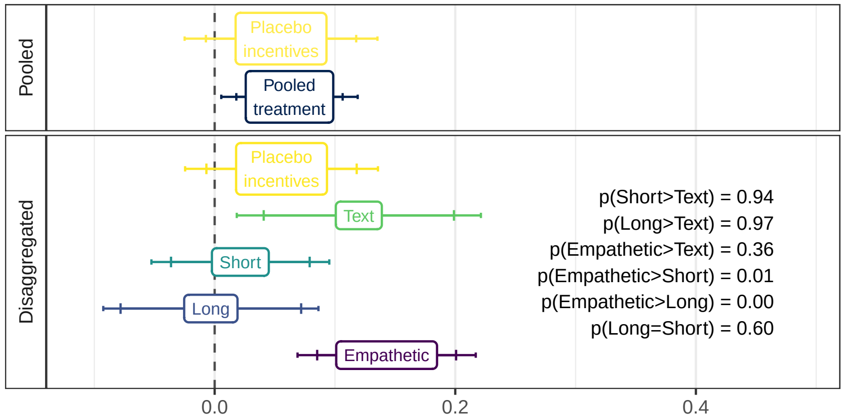
However, indicative of the challenges of generating organic demand for corrective information, the treatments combined with placebo quiz incentives resulted in significantly smaller increases in self-reported engagement, knowledge of fact-checks, and intended future take-up. Our results mirror prior findings suggesting that modest 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, although our findings suggest doing so is possible and could engender enduring engagement. 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 strongly with their assigned treatments.

The lower panel within each subfigure indicates that treatment take-up was fairly uniform across treatment conditions where participants were assigned to fact-check quiz incentives. There are no differences between these conditions in self-reported podcast listening in Figure 3a or in intended future take-up in Figure 3c. We do find that participants assigned to the *Empathetic* condition were more accurate in answering questions about recent fact-checks at endline than the other treatment conditions. However, this might reflect that empathetic content potentially increased users' information internalization. Regardless, the magnitude of this difference is relatively small. 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.

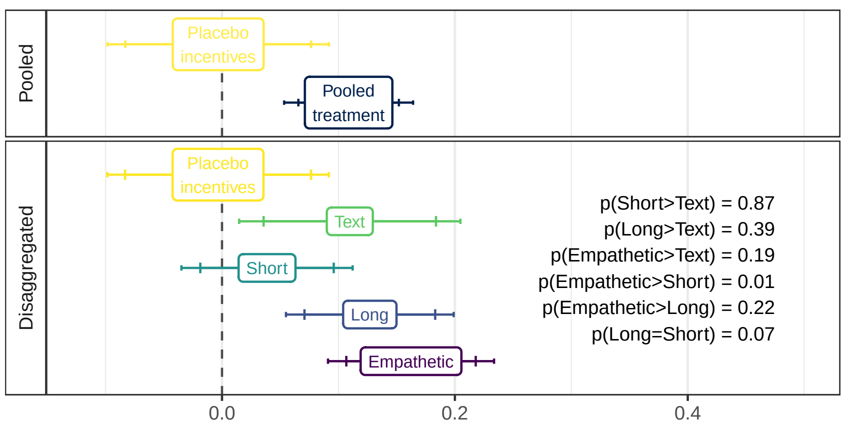
5.2 Discerning fact from fiction

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*. 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, and asked respondents to indicate how likely they believed each to be true. Figure 4a'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. Appendix Figures D1a and D1b further show that improved discernment is driven by respondents' greater distrust of false statements than greater trust of true statements. As the treatment variant tests in the lower panel illustrate, the pooled treatment effect is driven by the *Text* and *Empathetic* podcast conditions.

Second, we presented participants with four widespread conspiracy theories *not* investigated by Africa Check and asked respondents to indicate how likely each is to be true. The upper panel of Figure 4b indicates that any treatment with incentives to consume the fact-check quiz increased respondents' skepticism of conspiracy theories by 0.1 standard deviations, or an average of 0.12



(a) Discernment between true and fake news stories



(b) Identification of conspiracy theories

Figure 4: Treatment effects on discernment between fake and true news and belief in conspiracy theories

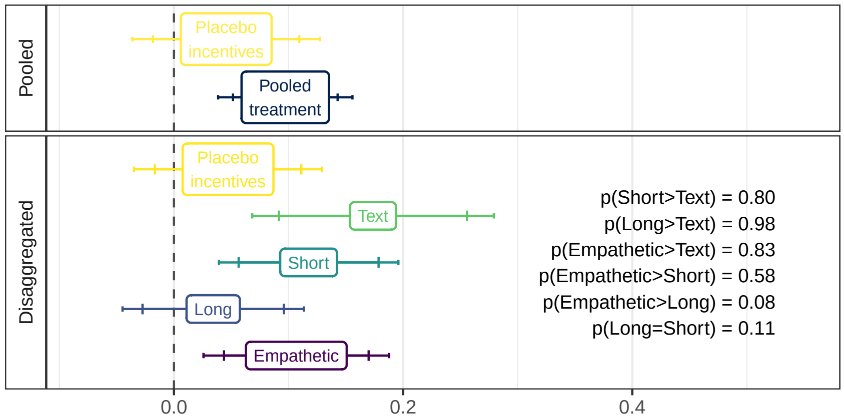
Notes: All outcomes are standardized inverse covariance-weighted indexes (see Table C1): (a): level of confidence in truthful claims and lack of confidence in false claims about how COVID spreads (true), whether matriculation exam scores inflated (false), if alcohol worsens infections (true), and that most workers are immigrants (false); (b) perceived likelihood that AIDS was intentionally created, Mandela died in 1985, COVID-19 vaccines have microchips, and vaccines used to reduce population. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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. Regression tables including all index components provided in Appendix Tables F4 and F5. Regression table of ICW indexes including all LASSO-selected controls provided in Appendix Tables F14-F15.

units on a five-point scale ($p < 0.01$). Increased discernment 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. Combined with participants' ability to distinguish true from false stories, sustained exposure to fact-checks reduced participants' susceptibility to fake news beyond the narrow content of the fact-checks. This suggests that sustained exposure to fact-checks can inoculate individuals against misinformation more broadly.

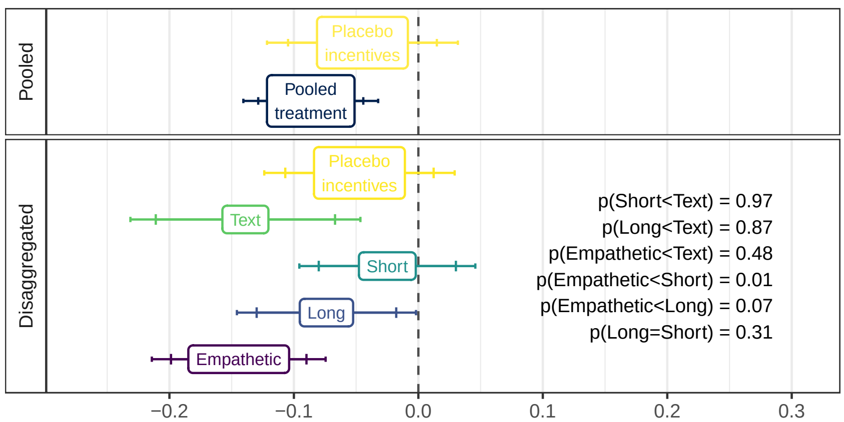
We next consider whether such generalized discernment is driven by the broader lessons imparted by Africa Check's fact-checking practices. Suggesting that prebunking is an important component of fact-checks, the upper panel of Figure 5a shows that repeated exposure to fact-checks led respondents to score 0.1 standard deviations higher on our information verification knowledge index ($p < 0.01$), which aggregates 13 items capturing good and bad practices for verifying news. Appendix Table F6 disaggregates the index, showing this effect principally reflects increases in respondents' 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 5a shows that the *Text*, *Short*, and *Empathetic* podcast modes of delivery were notably more effective ($p < 0.01$, $p < 0.01$, and $p < 0.05$, respectively) than the *Long* podcast.

Finally, effective inoculation might also reflect greater skepticism of platforms that supply a significant share of misinformation. Aggregating respondents' assessments of truth content on and trust in social media platforms (other than WhatsApp, through which fact-checks were delivered), the upper panel of Figure 5b shows that the treatments incentivizing participants to consume fact-checks reduced trust in social media platforms by 0.09 standard deviations ($p < 0.01$).[12] The effect is driven by each component of the index; for example, treatment reduced the share of respondents believing that social media information sources are credible by 17% ($p < 0.01$). In line with our previous results, the lower panel shows the largest effects for the *Text* and *Empathetic* podcast delivery formats ($p < 0.01$ and $p < 0.05$, respectively).

[12] Figure E3b shows that trust in information from close ties, including information sent from WhatsApp, modestly decreases.



(a) Knowledge of verification methods



(b) Trust in social media (besides WhatsApp)

Figure 5: Treatment effects on news verification knowledge and attitudes towards social media (besides WhatsApp)

Notes: All outcomes are standardized inverse covariance-weighted indexes (see Table C1): (a): separate indicators for correctly identifying 2 ways to avoid being misled, correctly identifying 7 methods to verify information, and correctly identifying 4 strategies fact-checkers use to verify information; (b) believes information from social media likely to be true, trusts information on social media, and thinks information on social media is most trustworthy. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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. Regression tables including all index components provided in Appendix Tables F6 and F8. Regression table of ICW indexes including all LASSO-selected controls provided in Appendix Tables F14-F15.

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 discern misinformation, verify suspicious information, and generally doubt content on social media *upon exposure*. Further, the heterogeneity across treatment groups, which all received fact-check quiz incentives and experienced similar effects on fact-check consumption, suggests that the main treatments—rather than differential consumption or the quizzes themselves—were responsible for the observed pattern of effects. In Appendix Figures D2a and D2b, we show no effects on participants' perception that misinformation is an important problem or that verification is important, nor any changes in their perception about the ease of fact-checking. This suggests that treated individuals became more *capable* of discerning fact from fiction, but not more motivated to do so.

5.3 Information consumption, verification, and sharing

Moving beyond efforts to inoculate participants *upon exposure* to misinformation, we assess 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 social media consumption (besides WhatsApp). Across the pooled and disaggregated estimations, Figure 6a reports substantively small and consistently statistically insignificant treatment effects. Furthermore, Appendix Figure E4 shows that consumption of news from traditional media and close 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 overall. Given that social media are consumed for many purposes beyond acquiring news, this illustrates the supply-side challenge of limiting misinformation exposure.

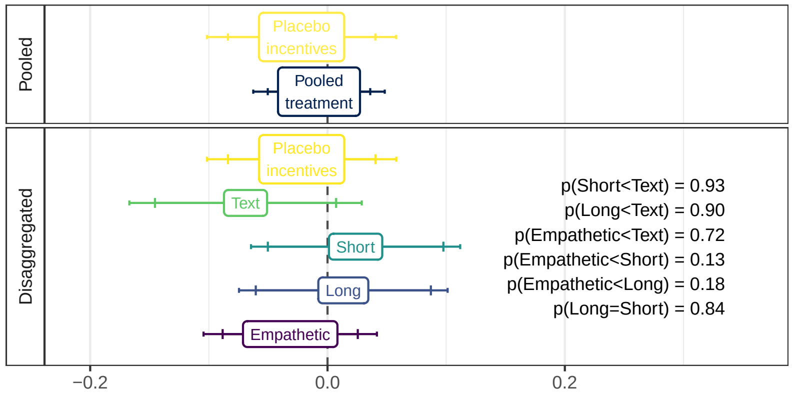
We similarly observe limited effects on respondents' active efforts to verify the truth of claims encountered outside the study. Specifically, Figure 6b shows that we fail to detect an increase in how often respondents reported trying to actively verify information they received through social media. Appendix Figure D3 indicates that, while verification 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 5a, 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 lower propensity to share suspected misinformation. The lower panel of Figure 6c shows that these participants became around 0.1 standard deviations less likely to report sharing information received via social media ($p < 0.05$), or a 0.1 unit reduction on our five-point scale capturing the frequency with which respondents share 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.

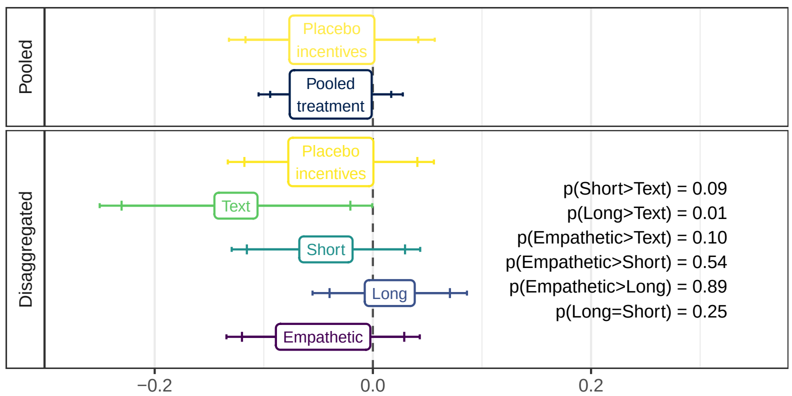
5.4 Attitudes and behaviors relating to COVID-19 and government

We finally turn to 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. Health-related misinformation, by emphasizing false cures or casting doubt on the severity of the pandemic, 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.

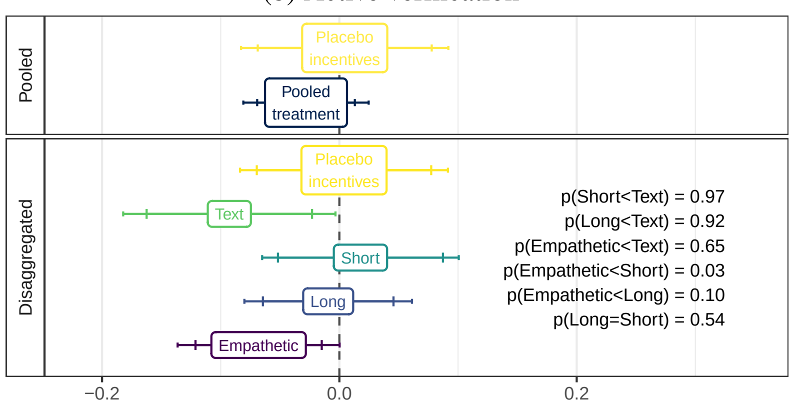
For our final set of outcomes, we therefore evaluate effects on indexes of attitudes and self-reported behaviors relating to COVID-19 and politics to assess whether the treatment mitigated the broader negative downstream consequences typically associated with exposure to misinformation.



(a) Social media consumption



(b) Active verification



(c) Sharing

Figure 6: Treatment effects on information consumption, verification and sharing

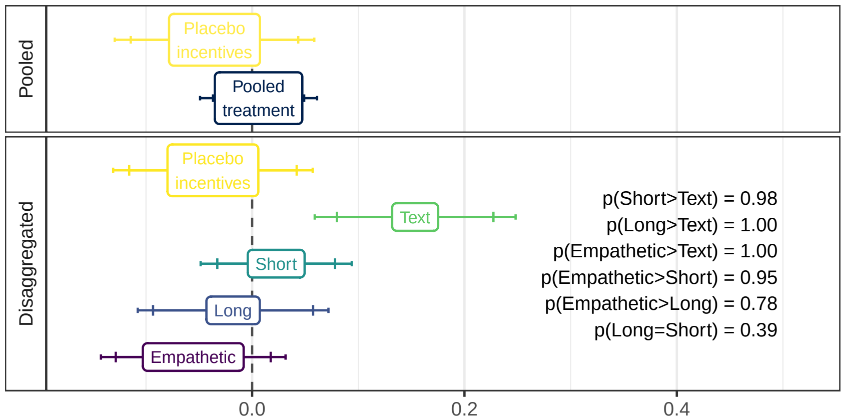
Notes: All outcomes are standardized (see Table C1): (a): how often gets news from non-WhatsApp social media; (b) how often actively verifies information; (c) how often shares stories on social media. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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. Regression tables including all components provided in Appendix Tables F9-F11. Regression table of ICW indexes including all LASSO-selected controls provided in Appendix Tables F14-F15.

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.

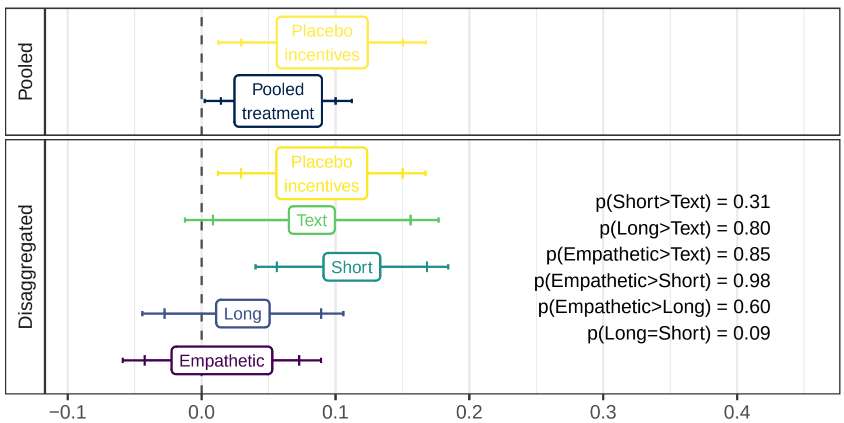
Overall, we detect modest effects after six months of exposure to fact-checks on such beliefs and behaviors. Figure 7a generally reports no treatment effect on COVID-19 beliefs and preventative behavior for the three podcast treatments with fact-check quiz incentives. However, we find that fact-checks delivered by short and simple text messages increased an index of health-conscious outcomes associated with COVID-19 by 0.14 standard deviations ($p < 0.01$). Particularly encouragingly, Appendix Table F12 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.

Figure 7b 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 ($p < 0.1$) is largely driven by the *Text* format—although the coefficient is not quite statistically significant ($p = 0.11$)—and *Short* podcast format ($p < 0.05$). Appendix Table F13 shows that these effects are primarily driven by significant increases in the extent to which respondents trusted information from politicians and the government.

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-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 of highly polarizing issues.



(a) COVID-19 beliefs and preventative behavior



(b) Views and attitudes about the government

Figure 7: Treatment effects on COVID-19 beliefs and preventative, and views and attitudes about the government

Notes: All outcomes are standardized inverse covariance-weighted indexes (see Table C1): (a): how many days stayed home in the last week, how many days visited other people indoors in the last week (reversed), how many days wore a mask in the last week, believes COVID-19 is a hoax (reversed), thinks lockdowns are necessary, trusts vaccines, and would get vaccinated; (b) central government performance appraisal, believes government handled COVID-19 well, faith in truth of information from politicians, trusts government/politicians most for information, level of trust in information from politicians, and would vote for regional incumbent party. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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. Regression tables including all components provided in Appendix Tables F12 and F13. Regression table of ICW indexes including all LASSO-selected controls provided in Appendix Tables F14-F15.

6 Conclusion

Misinformation on social media, due to its potentially negative consequences for political and health-related behaviors, is a growing concern around the globe. Misinformation has been linked to eroding trust in democratic institutions and political polarization; in South Africa, recent widespread misinformation has exacerbated racial tensions, fueled conflicts by stoking anger and fear, and substantially increased vaccine hesitancy.

While recent studies have advanced our understanding of how to mitigate the consumption of, and susceptibility to, misinformation online, most struggle to explain how sustained changes in beliefs and behaviors can be achieved outside controlled research environments. 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 organic *consumption* of corrective information, and how to induce *internalization* of the lessons imparted by fact-check content. The comparatively naturalistic setting of our intervention, along with its length, allowed us to examine whether fact-checking can play both *debunking* and *prebunking* roles by both correcting existing misinformation and warning participants about future misinformation. Our partnership with an existing fact-checking organization, Africa Check, 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 from the very beginning, an initial push towards consumption may subsequently activate latent demand. Policymakers should therefore target increasing the dissemination of fact-checked information in tandem with efforts to increase the public's appetite for such information. Our findings suggest that getting citizens over the initial hump could yield significant improvements in media literacy.

Second, while treated participants did not report altering behaviors that limit exposure to misinformation or active verification efforts, the robust effects on participants' capacity to discern fact from fiction—and willingness to act on this by not sharing unverified online content—indicate that the intervention contributed to participants' inoculation against misinformation *upon exposure*. 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. Different types of interventions, perhaps addressing access or production incentives in the broader media environment or consumption patterns within social networks, may be required to alter broader social media consumption patterns. In contrast, efforts to reduce exposure to misinformation could be more effectively targeted at its production than its consumption.

Third, not all treatment arms performed equally: the simple text-only treatment and empathetic podcast treatments were consistently the most effective delivery mechanisms for *internalization*. 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 modestly 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 key concerns of the social media age, reduced state capacity and declining faith in government.

References

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

Agunwa, Nkemakonam and Temiloluwa Alalade. 2022. "Dangers of gendered disinformation in African elections." WITNESS.
URL: 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.
URL: 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):327–342.

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):1481–1495.

Argote Tironi, 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):1325–1341.

Banerjee, Abhijit, Eliana La Ferrara and Victor H. Orozco-Olvera. 2019. "The Entertaining Way to Behavioral Change: Fighting HIV with MTV." NBER working paper.

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* 182:104–123.

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):946–959.

Berlinski, Nicolas, Margaret Doyle, Andrew M. Guess, Gabrielle Levy, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan and Jason Reifler. 2021. "The effects of unsubstantiated claims of voter fraud on confidence in elections." *Journal of Experimental Political Science* pp. 1–16.

Blair, Robert A., Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote and Charlene J. Stainfield. 2023. "Interventions to Counter Misinformation: Lessons from the Global North and Applications to the Global South." https://pdf.usaid.gov/pdf_docs/PA0215JW.pdf.

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* .

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.

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):236–243.

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):1531–1546.

Chen, Yuyu and David Y. Yang. 2019. "The Impact of Media Censorship: 1984 or Brave New World?" *American Economic Review* 109(6):2294–2332.

Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou and Brendan Nyhan. 2020. "Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media." *Political behavior* 42:1073–1095.

Cook, John. 2013. Inoculation Theory. In *The Sage Handbook of Persuasion: Developments in Theory and Practice*, ed. James Price Dillard and Lijiang Shen. Thousand Oaks, CA: SAGE Publications pp. 220–236.

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):127–150.

Gentzkow, Matthew, Michael B. Wong and Allen T. Zhang. 2021. "Ideological Bias and Trust in Information Sources.".

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):1–23.

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

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):15536–15545.

Guess, Andrew M., Nyhan Brendan and Jason Reifler. 2020. "Exposure to untrustworthy websites in the 2016 US election." *Nature Human Behaviour* 4(5):472–480.

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):110–126.

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):315–320.

International Federation of Journalists. 2021. "South Africa: Disinformation is the biggest threat to any election process.".
URL: `https://www.ifj.org/media-centre/news/detail/category/africa/article/south-africa-disinformation-is-the-biggest-threat-to-any-election-process`

Jerit, Jennifer and Yangzi Zhao. 2020. "Political Misinformation." *Annual Review of Political Science* 23(1):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):410–425.

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):106–131.

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.
URL: https://scholar.harvard.edu/files/jmarshall/files/tuning_in_voting_out_v6.pdf

Martel, Cameron, Gordon Pennycook and David G. Rand. 2020. "Reliance on emotion promotes belief in fake news." *Cognitive Research: Principles and Implications* 5(47).

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.

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

News24. 2019. "Fake news about xenophobia on social media aimed at ruining brand SA.".
URL: https://www.news24.com/news24/fake-news-about-xenophobia-on-social-media-aimed-at-ruining-brand-sa-govt-20190403

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* 42:939–960.

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. 2022. "Battling the Coronavirus Infodemic Among Social Media Users in Africa." *arXiv preprint arXiv:2212.13638* .

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* 592:590–595.

Pereira, Frederico Batista, Natalia Bueno, Felipe Nunes and Nara Pavao. forthcoming. "Inoculation Reduces Misinformation: Experimental Evidence from a Multidimensional Intervention in Brazil." *Journal of Experimental Political Science* .

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):133–147.

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.

Prior, Markus. 2007. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Quartz Africa. 2020. "WhatsApp is a key source of Covid-19 information for Africans." Quartz Africa.
URL: https://qz.com/africa/1871683/whatsapp-is-a-key-source-of-covid-19-information-for-africans

Reuters Institute. 2021. "Reporting elections: The frontline of the disinformation war." Reuters Institute for the Study of Journalism.
URL: https://reutersinstitute.politics.ox.ac.uk/news/reporting-elections-frontline-disinformation-war

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* .
URL: https://www.science.org/content/article/fighting-scientific-misinformation-south-african-perspective

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.
URL: https://doi.org/10.3390/vaccines10091379

Taber, Charles S. and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American Journal of Political Science* 50(3):755–769.

Tucker, Joshua A., Andrew M. Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social media, political polarization, and political disinformation: A review of the scientific literature." *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature* .

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):350–375.

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):686–693.

Wood, Thomas and Ethan Porter. 2019. "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence." *Political Behavior* 41:135–163.

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

Online Appendix

Sustaining Exposure to Fact-checks: Misinformation Discernment, Media Consumption, and its Political Implications

Table of Contents

A Methods **A1**

A.1 Recruitment and low-quality responses A1

A.2 Randomization A1

A.3 Financial incentives A1

A.4 Research ethics A2

A.5 Pre-specified hypotheses A2

A.6 Outcome measurement A3

A.7 Demand effects A4

B Examples of treatment **A5**

B.1 Examples of fact-checks A5

B.2 Examples of empathetic addition to podcast A6

B.3 Treatment delivery message primes A6

B.4 Examples of additional prime in delivery message A7

C Study design **A8**

C.1 Figures A8

C.2 Outcome variables A10

C.3 Balance and attrition A11

D Figures referenced in main text **A12**

E Figures referenced in supplementary materials and PAP **A14**

F Tables corresponding to figures in main text **A16**

G Pre-analysis Plan **A30**

A Methods

A.1 Recruitment and low-quality responses

To target a reasonably representative sample of the adult population of Facebook users in South Africa, recruitment ads on Facebook were stratified at the province-gender-age level, generating a total of 54 different ads that were targeted on the basis of the user's: (i) province (of which there are 9); (ii) gender; and (iii) age bracket (18-29, 30-49, or above 50 years old). Figure C1a provides an example of a recruitment ad, explaining that participants will receive airtime for participating in a social media study in South Africa.

Low-quality respondents were removed during the recruitment process using 3 attention-checking questions randomly appearing throughout the baseline survey. Questions were designed to be easy to respond to if respondents read the question somewhat carefully (e.g. "What year is it?"). We further restricted the sample to respondents who completed the baseline in more than eight minutes, which pilots of the baseline survey suggested was the minimum time required for the baseline survey to be comprehended and completed. Respondents who did not pass either check were excluded from randomization; consequently, dropped respondents are not correlated with treatment assignment. Their phone numbers were also prevented from restarting the baseline survey.

A.2 Randomization

We blocked-randomized individuals approximately once every two weeks by demographics, social media consumption, trust towards different news sources, and knowledge about misinformation. Figure 2 indicates the probabilities that participants were assigned to control and each treatment arm. 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. We used the `R` package `blocktools` to assign blocks, batch by batch, based on a greedy algorithm using Mahalanobis distance over seven predetermined baseline covariates. Our nested blocking strategy involved first creating blocks of size 38 (to ensure whole numbers of respondents were assigned across the various treatment combinations within a block) and then creating smaller sub-blocks of size 19 within each block. Our regression analyses use the blocks of size 38 rather than 19 because attrition often leaves the sub-blocks with missing treatment arms at endline. Whether we use the larger or smaller block fixed effects, results remain substantively unchanged.

A.3 Financial incentives

We administered small financial incentives (mobile airtime credits) to induce participation and continued engagement. Respondents who fulfilled all conditions for study enrollment (see above) received R30 (1.90 USD) in airtime. For each quiz, regardless of quiz type, respondents received R10 (0.62 USD) if they completed the quiz and an additional R10 if they answered a majority of the questions correctly. For a short midline survey, the results of which we do not report in the manuscript due to their broad similarity with the endline survey but with a much smaller set of outcomes, respondents were provided R30 (1.90 USD) for completion and an additional R10 if they answered a majority of the quiz questions embedded in the midline survey correctly. For the endline survey, respondents received R40 (2.50 USD) and an additional R10 if they answered a majority of the quiz questions embedded in the endline survey correctly. On average, endline respondents received a total of R155 (9.74 USD) through all components of the study. Figure C3a documents the share of participants completing each quiz during a given batch's study period, and the share of those completing each quiz who answered a majority of the questions correctly.

A.4 Research ethics

The design of our intervention reflected careful attention to the ethics of field experimentation and associated data collection consistent with APSA's *Principles and Guidance for Human Subjects Research* (2020).

First, with regard to the intervention itself, our expectation was that each treatment arm was likely to have positive effects on participants' ability to discern potentially harmful misinformation. This is because the interventions uniformly delivered misinformation-correcting information. While we preregistered theoretical expectations of *differences* between treatment arms in the magnitude of these positive effects, we did not anticipate—and, indeed, do not find—that any treatment arm would have effects consistent with harmful welfare consequences. At the same time, participants assigned to control were not prevented from independently signing up to receive fact-checking programming from Africa Check, outside of the confines of the study, if they desired to do so.

Second, with regard to participation and consent, we solicited informed consent from all participants in the study and did not use any sort of deceit relating to the study's purpose. Participants were free to take, or not take, the optional monthly quizzes as well as the subsequent surveys. While we did use financial incentives in the form of mobile airtime transfers (see Section A.3), these were relatively small overall and served as small incentives to maintain the engagement of participants through a relatively long study period overall. Participants were free to drop from the study at any time, all their responses were anonymized, and we anticipated that participants would face no retaliation or repercussions from taking part in the study.

Third, with regard to the broader impact of the study, we expected that the limited sample size of the participants involved would render any wider political consequences highly unlikely (beyond informing the programming strategy of the implementing partner). While we collaborated with Africa Check to implement the study, they had no ability to veto or review study conclusions prior to writing the paper and the authors have no conflicts of interest relating to the organization.

A.5 Pre-specified hypotheses

We preregistered the following hypotheses for pooled treatment effects, which correspond to the outcomes presented in the main text and in the top panel of each subfigure:

- **Treatment take-up:** Access to treatment increases both exposure to, and knowledge about, information covered by the treatment deliveries (H1).
- **Discerning fact from fiction:** Fact-check treatments would increase participants' capacity to identify, and express skepticism on the basis of, characteristics of misinformation (H6);

reduce trust in social media information (H3); and increase the perceived extent of misinformation on social media (H2).

- **Information consumption, verification, and sharing:** Fact-check treatments would decrease information consumption and sharing from social media (H4), increase awareness and attention paid to information on social media (H5), and increase active fact-checking behavior (H7).
- **COVID-19 and political attitudes and behavior:** Fact-check treatments would increase participants' knowledge and beliefs in the severity of COVID-19 and their willingness to take preventative measures (H9) and improve participants' perceptions of government performance (H8).

The corresponding hypothesis from our pre-analysis plan is noted in parentheses. Overall, we find evidence consistent with H1, H3, H4 (with regard to sharing), H6, H8, and H9. In addition to the pooled effects, we hypothesized that treatment would be more effective for incentivized ("fact-check quizzes") rather than unincentivized ("placebo quizzes") treatments, which we find strong support for. Between treatment arms, we hypothesized that (1) effects would be greater for podcasts rather than text messages, and (2) *Empathetic* podcasts rather than *Long* podcasts, but (3) we made no directional predictions for differences between the *Long* and *Short* podcasts. We find evidence consistent with (2) but not (1), since the text treatment was ultimately highly effective. Finally, we preregistered an expectation of greater treatment effects for treatments delivered using a social prime that highlighted the importance of fact-checking for social good, which we also found to be the case (see below).

A.6 Outcome measurement

All our main outcomes are inverse covariance weighted (ICW) indexes (see Anderson 2008). Each such outcome aggregates individual survey items in line with the families outlined in our pre-analysis plan, and is standardized with respect to the control group mean and standard deviation. Each grouping of outcomes contains several ICW outcome indexes capturing different types of outcome within the family. These groupings are provided in Table C1.

Missing responses were imputed as follows. "Don't know" responses to specific questions were coded as "negative" responses relative to the expected treatment effect sign, which were all normalized to positive; e.g. when the respondents were asked about listening to podcasts, "Don't know" is coded as "Never." Similarly for the importance of an issue, "Don't know" is coded as "Not at all important". In turn, when "Don't know" relates to a Likert scale, "Don't know" is coded as the median/neutral option (e.g. as "neither agree not disagree").

The final indexes we settled on largely conform with the indexes specified in the pre-analysis plan. However, we note below some deviations designed to focus attention on theoretically-relevant outcomes.

First, for exposure to the intervention, we examine podcast take-up and knowledge of the content of the podcast separately to distinguish self-reported attention from internalization; we cut an index item about the frequency with which participants report being alerted to fake news on social media because it was originally designed to test a distinct mechanism proposed in the literature

(Pennycook et al. 2021), but we found limited support for it (see Figure E1). We further added future take-up as an additional indicator of treatment take-up once the small financial incentives to participate in the study had been removed.

Second, for trust in social media, the index focuses on Facebook, Instagram, and Twitter. We exclude WhatsApp because the fact-checking intervention was delivered via WhatsApp and hence results are difficult to interpret. Figure E3b shows that trust in information from close ties, including information sent by these ties from WhatsApp, modestly decreases. Third, for consumption of social media, we exclude WhatsApp for the same reason. We also examine the consumption and sharing of information separately to examine effects on both important outcomes.

Fourth, our discernment outcomes relating to conspiracy theories were not pre-registered, but provide a valuable check on citizen evaluations of claims that could be the subject of misinformation.

Fifth, we distinguish between active verification efforts and knowledge about the correct way to verify information. For active verification, we solely focus on the frequency with which a respondent reports fact-checking information (see Figure 6b and Table F10). We use the following variables for knowledge on how to verify: the perceived importance of fact-checking, verifying by seeking out dedicated fact-checkers, and levels of knowledge about how and where to check misinformation (see Figure 5a and Table F6). We exclude the variable on whether they share fact-checks with friends and family, as that does not fall appropriately into either active verification or knowledge of how to verify information (see Figure E2).

Finally, for attitudes toward the government, we deviate from the pre-analysis plan in three ways to focus on trust in and appraisals of government politicians and performance: (i) we add items relating to trust in government and politicians and the information they provide (see Figure 7b); (ii) we exclude two questions eliciting perceptions of government capacity (see Figure E5 for results) and two questions on populism-related beliefs (see Figure E6 for results), on the basis that these questions were worded to capture beliefs about how government *ought* to behave rather than concrete government appraisals.

A.7 Demand effects

Because our outcomes are derived from survey measures, participants who were assigned to treatment arms, in principle, may have responded to questions based on perceptions of what answers were more desirable. We provide evidence against social desirability bias in three ways.

First, social desirability bias is unlikely to account for differences across treatment arms. Consistent differences in treatment effects across the treatment arms suggest that particular components of the intervention did elicit real change in participants' knowledge and beliefs about information from online news media. This interpretation of our findings is bolstered by results from questions that test participants' capacity to discern true from false news and their ability to identify conspiracy theories. The information in these two sets of questions were *not* covered by the information Africa Check delivered weekly. These knowledge questions are difficult to falsify, as they require participants to be aware of current events and better adjudicate a piece of news' credibility. Moreover, treated participants were better able to recall treatment content and identify plausible verification methods—other outcomes that are less susceptible to social desirability bias.

Second, demand effects are unlikely to explain our set of results, which show differences between the intervention's success in increasing participants' knowledge and awareness versus actual behavioral change. If participants who were assigned to treatment arms selected socially desirable survey responses, we would expect participants to also report greater behavioral changes with respect to social media consumption and active verification of online content. Our findings indicate that this is not the case: estimated treatment effects suggest that actual behavior with respect to social media interaction is hard to shift despite consistent exposure to the intervention.

Third, we examine a behavioral outcome that is unlikely to be affected by social desirability bias. Every treatment delivery from Africa Check also included a message that encouraged participants to submit fact-checking requests to discern true participant interest in the fact-checking information. Participants could submit text or forward videos, pictures, or links to the Africa Check phone number for fact-checking. Estimates in Figure E7 show that treated participants were indeed more likely to submit fact-check requests. Importantly, the incentivized *Text* treatment participants were the most likely to send in fact-checking requests in comparison to all other treatment arms ($p < 0.01$). The particular effectiveness of the *Text* treatment, in comparison to the other treatment arms, is consistent with our other survey outcomes and assuages concerns about demand effects across the study.

B Examples of treatment

B.1 Examples of fact-checks

The fact-checks conducted by Africa Check's were deemed true, false, misleading, or uncertain (unsubstantiated). Figure 1) shows that these fact-checks covered (broadly) eight families of issues but often touch upon more than one set of issues. Below are examples of each type of issue:

- **Politics:** "Did a R200m Covid-19 vaccine tender go to the daughter of South African premier? This is incorrect!"
- **Economy:** "Beware of false job adverts for the South African police. It's a job scam."
- **Race/Xenophobia:** "Did a recent tweet by Julius Malema encourage attacks on 'racist farms'? No, it's fake!"
- **COVID-19:** "No, a World Health Organization head didn't say Covid vaccines kill kids."
- **Other Health:** "There is no scientific evidence that a mixture of bitter gourd leaves and snails is a remedy for stroke."
- **Crime:** "Has the murder rate for the North West nearly doubled from 2020 to 2021? Yes, but the Covid-19 lockdown skewed the comparison."
- **Society:** "Are there 5.6 billion women in the world to just 2.2 billion men? Nope, not even close!"
- **Miscellaneous fun facts:** "There is no elephant-shaped mountain in Oregon, US – the image that has been circulating was photoshopped by an artist."

B.2 Examples of empathetic addition to podcast

- "Misinformation about vaccine and vaccine mandates can be scary. Especially when it suggests that we may be forced to do something or the vaccines could have side effects. So it's really important that we check claims like this before we pass them on."
- "With the rising number of daily COVID-19 positive cases and of course the new variant, many people may be feeling anxious about an onset of cold or flu symptoms. Even seasonal allergies. And the panic around this may lead you to fall for misinformation on how to mitigate symptoms as well as unverified remedies on how to get better quicker. Which is the case with this claim."
- "You may have seen pictures or videos shared on social media of gas or paraffin heater incidents that led to serious burn-related injuries. And this first claim may make you feel anxious or fear for the safety of your friends or family members who regularly use these appliances. And you might want to share safety hacks to protect your loved ones and to caution them to take extra care to avoid danger with appliances this winter. But sometimes, these aren't entirely true..."

B.3 Treatment delivery message primes

All treatment arms included a short message that accompanied the delivery of the treatment. Within each treatment arm, a random half of the participants received a message that simply introduced the fact-check information being delivered (*Factual*), while the other half received a message that primed participants about the information's importance to encourage consumption of the fact-check material (*Prime*). We expected treatment effects to be particularly concentrated among participants assigned to *Prime* rather than *Factual* messages.

For our main analysis, we focus on the preregistered approach of pooling the *Factual* and *Prime* messages within each form of treatment. We now examine potential complementarities between these treatments and the *Prime* message. We return to examine the outcomes for which *Text* and all podcast treatments produced significant impacts: discernment between fake and true information; identification of conspiracy theories; and verification knowledge. The variation in treatment delivery message does not induce clear differential effects on our other outcomes.

The message priming the social importance of misinformation increased discernment (results omitted due to length constraints and available upon request). Across two treatment arms—*Text* and *Empathetic* podcast paired with *Fact-check* quizzes—we find that messages with the social *Prime* significantly increased the likelihood that participants were able to discern between fake and true information. While the incentivized *Long* podcast also performed better when paired with a *Prime* message, the treatment combination is not statistically distinguishable from the *Control* condition. We similarly find that the *Prime* message amplified the impact of other treatments on the likelihood of doubting conspiracy theories. When primed, participants were more likely to identify conspiracy theories across three incentivized treatment arms: the *Text* treatment, the *Long* podcast, and the *Empathetic podcast*. Moreover, the *Prime* message—when paired with the incentivized *Text*, *Short* podcast, and *Empathetic* podcast—was once again significantly more likely to help participants identify correct strategies for verifying information.

Overall, we find evidence consistent with the inclusion of a *Prime* message when encouraging participants to internalize their assigned treatments—particularly for the incentivized *Text* and *Empathetic* podcasts. These originally identified effects are then amplified by a *Prime* message which repeatedly reminded participants of fact-checking's importance. Because the prime did not increase reported *consumption* but did increase knowledge about its content, the results are primarily driven by participants' *internalization* upon exposure.

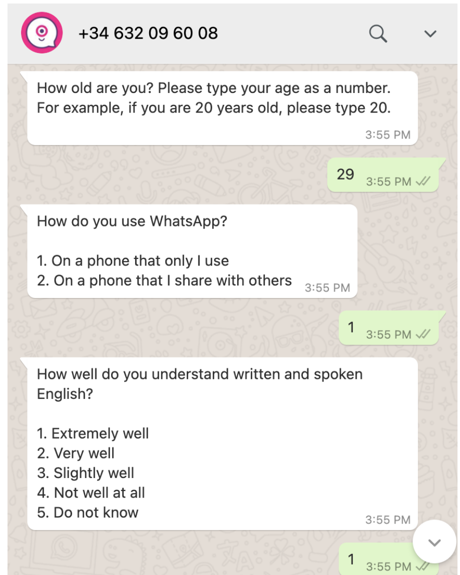
B.4 Examples of additional prime in delivery message

- "Myth busters and fake news debunkers play a vital role in checking the facts online! Here are the facts about three viral online messages so you can prevent your friends and family from being fooled by false information."
- "False information can be dangerous. Sometimes it can be deadly. Play your part in sharing accurate information online to help protect your friends and family. Here are the facts about three viral online messages:"
- "False and misleading information can be dangerous. When it comes to health issues, it can be deadly. Verify before you share message online to keep your fiends and family safe. They'll thank you for it! We've fact-checked three viral messages for you:"

C Study design

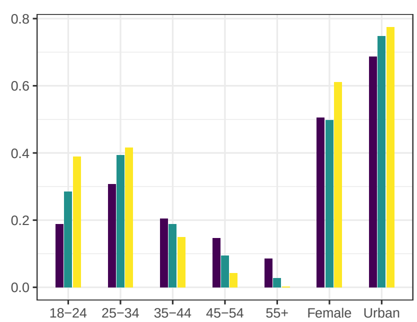
C.1 Figures

(a) Recruitment Facebook ad

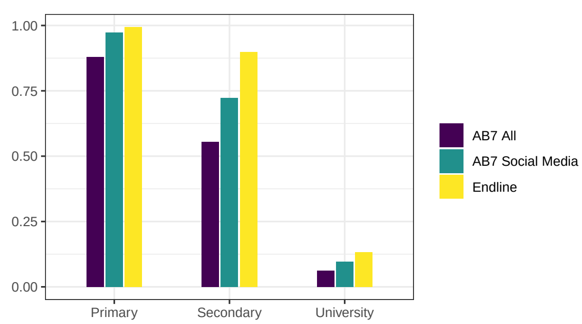


(b) Survey through WhatsApp chatbot

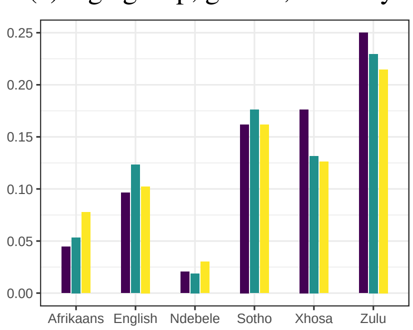
Figure C1: Recruitment and surveying



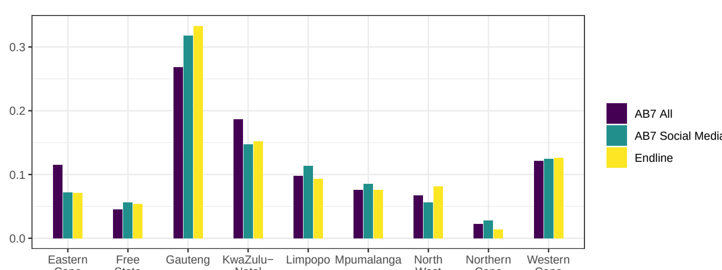
(a) Age group, gender, urbanity



(b) Education

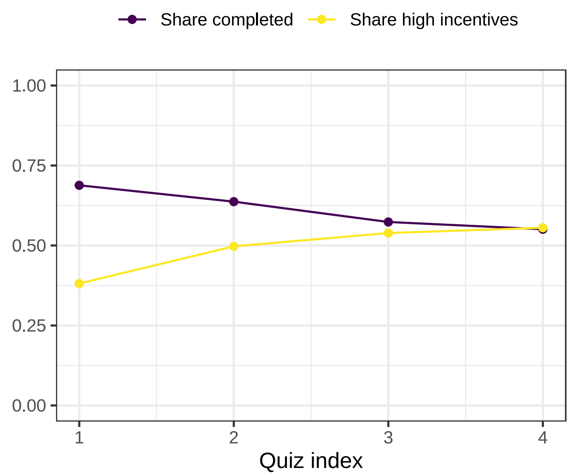


(c) Ethnicity

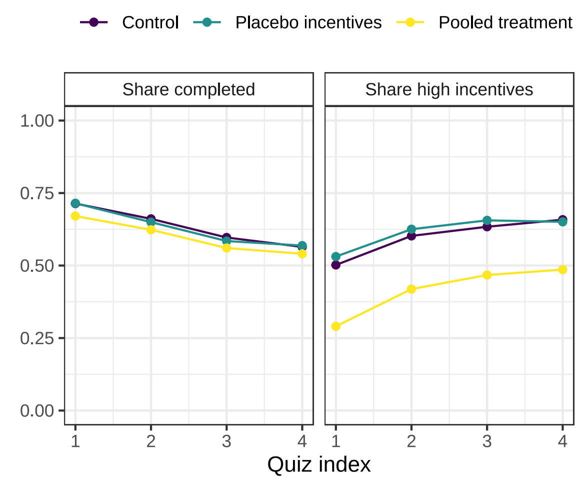


(d) Province

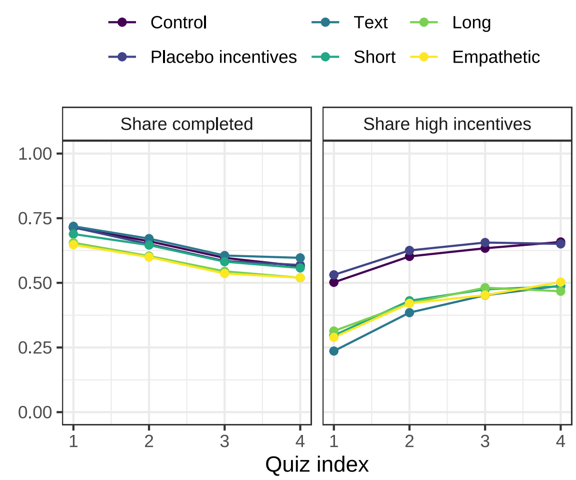
Figure C2: Comparison of endline sample with Afrobarometer round 7 (2018)



(a) Quiz engagement and incentive payments (overall)



(b) Quiz engagement and incentive payments (pooled treatment)



(c) Quiz engagement and incentive payments (disaggregated treatment)

Figure C3: Quiz engagement over study

Notes: Figure plots average participation, and average share of participants answering more than 50% of questions correctly, through study quizzes (fact-check or placebo) between baseline and endline.

C.2 Outcome variables

Table C1: Outcome variables

Outcome variable	Variable definitions	Mean	SD	Range
Treatment take-up				
Podcast take-up (Fig. 3a)	How often listen to podcasts	3.24	1.25	[1,5]
	Included “What’s Crap on WhatsApp” in selection of podcasts they listened to	0.41	0.49	[0,1]
Treatment knowledge (Fig. 3b)	Number of correct responses from 6 questions on fact-checked content	2.75	1.56	[0,6]
Future take-up (Fig. 3c)	Want vaccine info from Africa Check	0.72	0.45	[0,1]
	Want Africa Check’s fact checking content	0.85	0.36	[0,1]
	Want Africa Check reminders to pay attention to misinformation	0.71	0.45	[0,1]
	Stay subscribed (or start subscribing) to “What’s Crap on WhatsApp”	0.83	0.37	[0,1]
Discerning fact from fiction				
Discernment between T/F news (Fig. 4a)	How COVID-19 spreads (true)	4.45	0.91	[1,5]
	Matriculation scores to be inflated (false) (-)	3.11	1.34	[1,5]
	Alcohol decreases ability to fight infections (true)	3.51	1.27	[1,5]
	Almost 100% of workers in SA are foreign (false) (-)	2.89	1.31	[1,5]
Identification of conspiracy theories (Fig. 4b)	Not at all likely to very likely: AIDs intentionally created	3.69	1.37	[1,5]
	Not at all likely to very likely: Nelson Mandela died in 1985	3.82	1.38	[1,5]
	Not at all likely to very likely: COVID-19 vaccines used to implant chips	3.70	1.34	[1,5]
	Not at all likely to very likely: Vaccines used to reduce world’s population	3.72	1.34	[1,5]
Verification knowledge and trust				
Knowledge of verification methods (Fig. 5a)	To avoid being misled: Seek info from reputable org	0.36	0.48	[0,1]
	To avoid being misled: Ask other people to avoid being misled (-)	0.13	0.34	[0,1]
	To verify: Ask people I know in person (-)	0.71	0.46	[0,1]
	To verify: Ask people I know through WhatsApp (-)	0.82	0.39	[0,1]
	To verify: Ask people I don’t know well on WhatsApp group (-)	0.91	0.29	[0,1]
	To verify: Go to fact-checker	0.49	0.50	[0,1]
	To verify: Submit a fact-checker request	0.21	0.40	[0,1]
	To verify: Ask people I know by posting on social media (-)	0.87	0.33	[0,1]
	To verify: Use the internet to fact-check	0.46	0.50	[0,1]
	Verify strategies: Ask experts	0.42	0.49	[0,1]
	Verify strategies: Check source popularity (-)	0.63	0.48	[0,1]
	Verify strategies: Use reverse image searches	0.16	0.36	[0,1]
	Verify strategies: Talk to others (-)	0.82	0.38	[0,1]
Trust in social media besides WhatsApp (Fig. 5b)	Likely to be true: Information from other social media (FB, Twitter, Instagram)	2.83	0.75	[1,5]
	Trust: Information received from other social media (FB, Twitter, Instagram)	2.88	1.04	[1,5]
	Trust the most for information: Other social media (FB, Twitter, Instagram)	0.16	0.37	[0,1]
Info consumption, verification, and sharing				
Online and social media consumption (Fig. 6a)	Go to source for news: other social media (Facebook, Twitter, Instagram)	0.42	0.49	[0,1]
Verification (Fig. 6b)	How often verify information seen on social media	3.83	1.10	[1,5]
Sharing (Fig. 6c)	How often share social media info shared by others	2.83	1.11	[1,5]
COVID-19 and political attitudes				
COVID-19 beliefs and behaviors (Fig. 7a)	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]
	Strongly disagree to strongly agree: COVID-19 is a fake disease (-)	4.36	1.11	[1,5]
	Definitely to definitely not: COVID-19 lockdown justified (-)	3.21	0.92	[1,4]
	Strongly disagree to strongly agree: Would take available vaccine	3.49	1.54	[1,5]
	Strongly distrust to strongly trust: COVID-19 vaccines in South Africa are safe	3.89	1.37	[1,5]
Views and attitudes about government (Fig. 7b)	Trust information from politicians and gov officials	2.89	1.20	[1,5]
	Most trustworthy sources: Selected “Government officials”	0.30	0.46	[0,1]
	Most trustworthy sources: Selected “Politicians and other public figures”	0.13	0.34	[0,1]
	How likely information from politicians and gov officials are true	3.02	0.95	[1,5]
	Vote for regional incumbent (vote tomorrow in parl elections: ANC, DA, EFF, IFP, VF+)	0.23	0.42	[0,1]
	Very badly to very well: National government’s general performance	2.38	1.20	[1,5]
	Very badly to very well: National government handling COVID-19 crisis	3.09	1.22	[1,5]

Descriptive statistics for all variables used in figures in main paper.
Variables indicated with (-) indicate that variable has been reversed for use in index before providing summary statistics.

C.3 Balance and attrition

Table C2: Attrition

	Attrition	
	(1)	(2)
A. Pooled estimation		
Placebo incentives	0.023	0.021
	(0.017)	(0.016)
	[0.172]	[0.209]
Pooled treatment	-0.014	-0.017
	(0.012)	(0.012)
	[0.220]	[0.137]
B. Disaggregated estimation		
Placebo incentives	0.023	0.021
	(0.017)	(0.017)
	[0.171]	[0.197]
Text information	-0.022	-0.026
	(0.021)	(0.021)
	[0.302]	[0.215]
Short podcast	0.002	-0.003
	(0.016)	(0.015)
	[0.878]	[0.846]
Long podcast	-0.021	-0.022
	(0.015)	(0.015)
	[0.172]	[0.156]
Empathetic podcast	-0.021	-0.022
	(0.016)	(0.015)
	[0.171]	[0.145]
Controls	×	✓
Directional hypothesis	×	×
Control Mean	0.51	0.51
Control SD	0.50	0.50
R^2	0.12	0.16
Observations	8947	8947

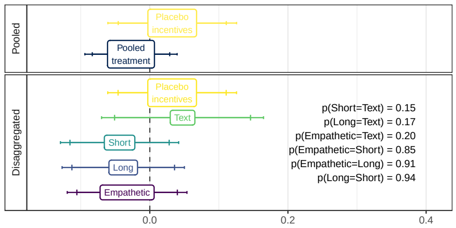
Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets.

Table C3: Balance on pre-treatment outcomes

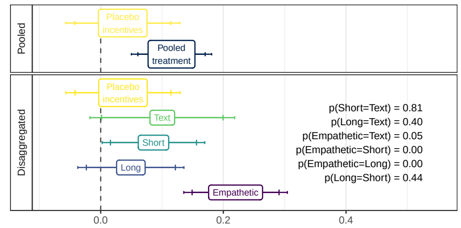
Variable	$p(\tau_{pooled} = 0)$	$p(\tau_{disagg.} = 0)$
A. Socio-demographic		
Gender: Female	[0.990]	[0.666]
Locality: Urban	[0.573]	[0.297]
Locality: Peri-urban	[0.572]	[0.909]
Locality: Rural	[0.558]	[0.796]
Age: 18-24	[0.791]	[0.620]
Age: 25-34	[0.176]	[0.463]
Age: 35-44	[0.518]	[0.761]
Age: 45-54	[0.147]	[0.095]
Age: 55+	[0.371]	[0.441]
Education: Primary	[0.495]	[0.204]
Education: Secondary	[0.857]	[0.744]
Education: University	[0.790]	[0.707]
Province: Eastern Cape	[0.328]	[0.643]
Province: Free State	[0.629]	[0.898]
Province: Gauteng	[0.870]	[0.994]
Province: KwaZulu-Natal	[0.796]	[0.388]
Province: Limpopo	[0.956]	[0.512]
Province: Mpumalanga	[0.499]	[0.138]
Province: Northern Cape	[0.032]	[0.204]
Province: North West	[0.271]	[0.664]
Province: Western Cape	[0.493]	[0.879]
B. Baseline survey responses		
Verify challenge	[0.430]	[0.783]
Consume close friends	[0.784]	[0.917]
Consume social media	[0.190]	[0.426]
Consume traditional media	[0.257]	[0.345]
Consume WhatsApp	[0.409]	[0.834]
COVID-19 beliefs and behavior	[0.159]	[0.465]
Podcast take-up	[0.877]	[0.905]
First stage placebo	[0.609]	[0.603]
Misinformation harmful	[0.878]	[0.501]
Sharing	[0.962]	[0.715]
Trust close friends	[0.663]	[0.806]
Trust social media	[0.482]	[0.747]
Trust organizations	[0.989]	[0.872]
Trust traditional media	[0.850]	[0.930]
Trust WhatsApp	[0.562]	[0.903]
Active verification	[0.722]	[0.179]
Verification knowledge	[0.161]	[0.271]

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects. $p(\tau_{pooled} = 0)$ provides the p-value from a test of joint significance of coefficients in the pooled estimation (control; placebo incentives; pooled treatment); $p(\tau_{disagg.} = 0)$ provides the p-value from a test of joint significance of coefficients in the disaggregated estimation (control; placebo incentives; text; short; long; empathetic).

D Figures referenced in main text



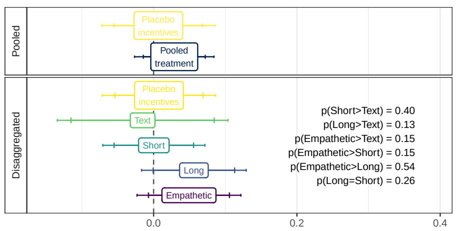
(a) Correct discernment of true news stories



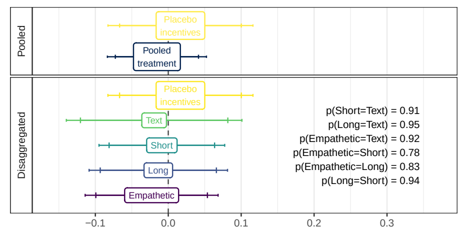
(b) Correct discernment of false news stories

Figure D1: Treatment effects on discernment between fake and true news

Notes: All outcomes are standardized inverse covariance-weighted indexes: (a): level of confidence in truthful claims about how COVID spreads (true) and if alcohol exacerbates infections (true); (b) lack of confidence in false claims about inflation of matriculation exam scores (false) and most workers being immigrants (false). Estimated using Equation (1); while the interior and exterior bars represent 90% and 95% confidence intervals.



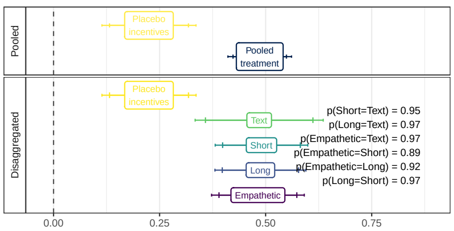
(a) Verification is important



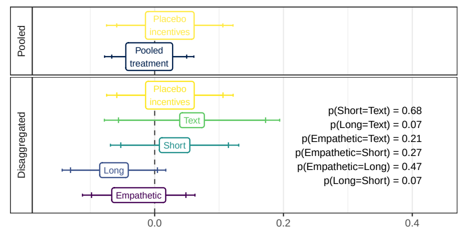
(b) Fact-checking is challenging

Figure D2: Treatment effects on verification and ease of fact-checking

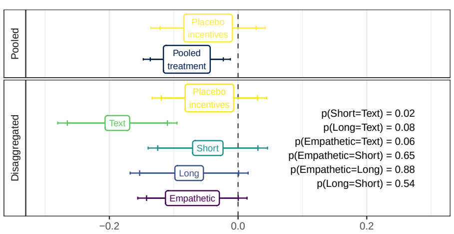
Notes: All outcomes are standardized inverse covariance-weighted indexes: (a): thinks it is important to verify information; (b): challenging to verify information due to knowledge, irrelevant fact-checks, distrust fact-checkers, too expensive, overwhelming information, takes too long. Estimated using Equation (1); while the interior and exterior bars represent 90% and 95% confidence intervals.



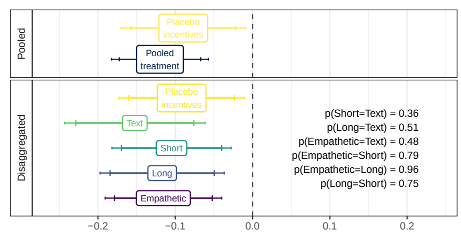
(a) Verify through Africa Check



(b) Verify through other fact-checkers



(c) Verify through online and social media



(d) Verify through traditional media

Figure D3: Treatment effects on the use of different information sources for verification

Notes: All outcomes are standardized inverse covariance-weighted indexes: (a): lists WCW as a source for fact-checking; (b) lists AFP or Snopes as a source; (c) lists Facebook, Google, Moya, Telegram, Twitter, WhatsApp, or YouTube as a source; (d) lists News24 or SABC as a source. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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.

E Figures referenced in supplementary materials and PAP

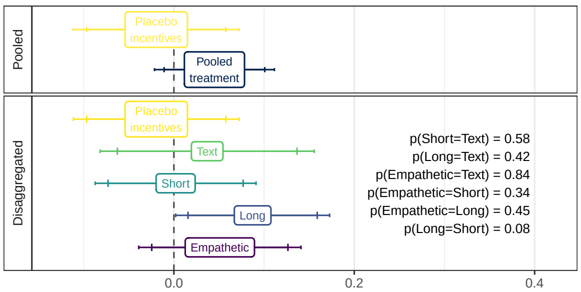


Figure E1: Being alerted about fake news

Notes: Outcome is standardized: How often participant is alerted about fake news. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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.

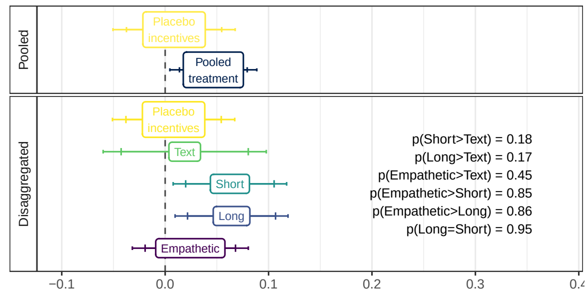
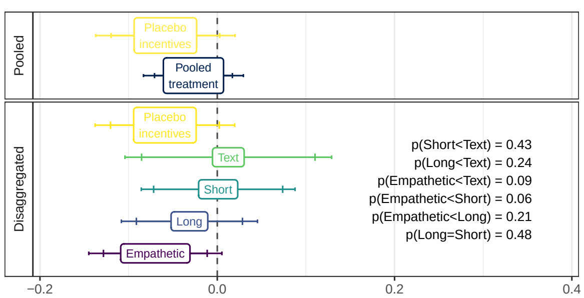
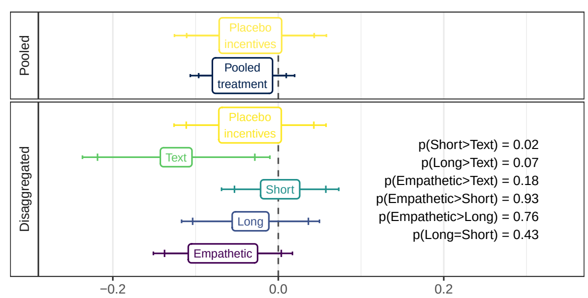


Figure E2: Alerting others about fake news

Notes: Outcome is standardized: How often participant reports alerting others about misinformation. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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.



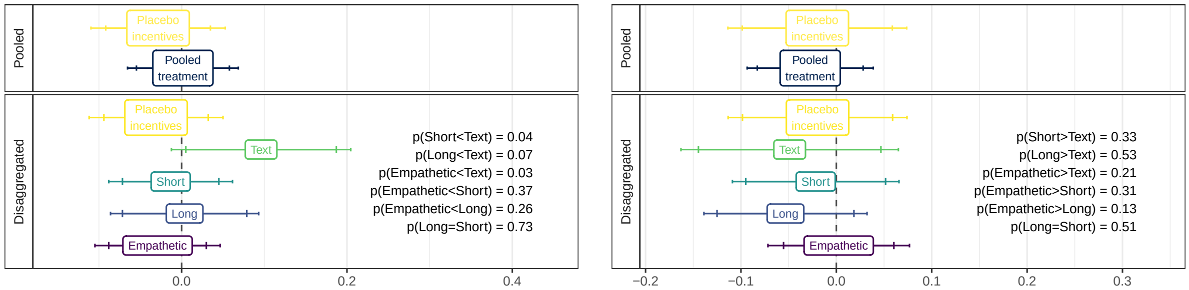
(a) Trust in traditional media



(b) Trust information sent by close ties

Figure E3: Treatment effects on trust in different sources

Notes: All outcomes are standardized inverse covariance-weighted indexes: (a): how true is info on radio/TV, trusts newspapers most for information, trusts information from radio/TV; (b) how true is info from friends and family, trusts info from friends and family, trusts WhatsApp messages from friends and family. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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.



(a) Traditional media consumption

(b) Consumption of news from close ties

Figure E4: Treatment effects on consumption from different sources

Notes: All outcomes are standardized inverse covariance-weighted indexes: (a): how often gets news from radio/TV; (b) how often gets news from friends and family. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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.

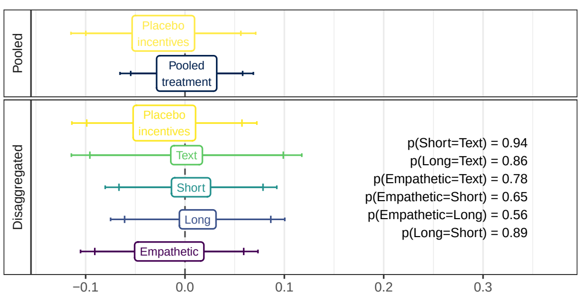


Figure E5: Perceptions of government capacity

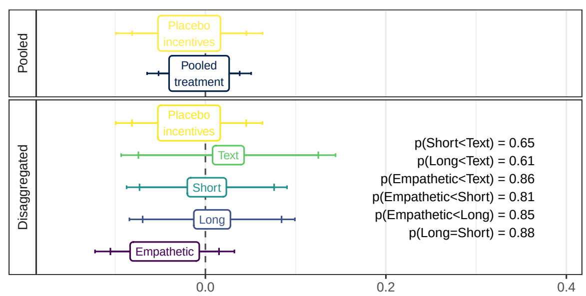


Figure E6: Populist attitudes

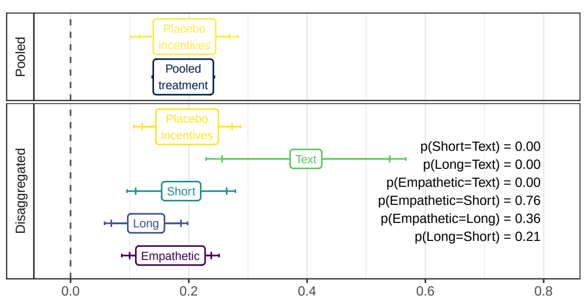


Figure E7: Fact-check requests

Notes: **Fig E5:** Outcome is standardized inverse covariance-weighted index comprising perception of government capacity to provide roads; perception of government capacity to supply electricity. **Fig E6:** Outcome is standardized inverse covariance-weighted index comprising perception of policies benefit elites; perception that ordinary people have no influence over policy. **Fig E7:** Outcome is a standardized indicator for participant submitting a fact-check request to Africa Check. Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with 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.

F Tables corresponding to figures in main text

Table F1: Podcast take-up

	ICW: Podcast take-up		How often listens to podcasts		Listens to WCW	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Pooled estimation						
Placebo incentives	0.416	0.424	0.018	0.023	0.247	0.251
	(0.054)	(0.054)	(0.059)	(0.059)	(0.025)	(0.024)
	[0.000]	[0.000]	[0.381]	[0.348]	[0.000]	[0.000]
Pooled podcast	0.651	0.646	0.132	0.123	0.361	0.360
	(0.036)	(0.035)	(0.041)	(0.041)	(0.015)	(0.015)
	[0.000]	[0.000]	[0.001]	[0.001]	[0.000]	[0.000]
B. Disaggregated estimation						
Placebo incentives	0.321	0.323	0.020	0.021	0.188	0.190
	(0.050)	(0.049)	(0.055)	(0.055)	(0.023)	(0.022)
	[0.000]	[0.000]	[0.355]	[0.354]	[0.000]	[0.000]
Text information	-0.020	-0.014	-0.088	-0.084	0.014	0.018
	(0.060)	(0.059)	(0.072)	(0.071)	(0.024)	(0.025)
	[0.744]	[0.818]	[0.224]	[0.232]	[0.282]	[0.232]
Short podcast	0.648	0.638	0.160	0.153	0.349	0.345
	(0.047)	(0.047)	(0.052)	(0.052)	(0.021)	(0.021)
	[0.000]	[0.000]	[0.001]	[0.002]	[0.000]	[0.000]
Long podcast	0.646	0.646	0.120	0.114	0.360	0.361
	(0.048)	(0.048)	(0.054)	(0.054)	(0.021)	(0.021)
	[0.000]	[0.000]	[0.013]	[0.017]	[0.000]	[0.000]
Empathetic podcast	0.665	0.656	0.116	0.099	0.375	0.374
	(0.048)	(0.047)	(0.054)	(0.053)	(0.021)	(0.021)
	[0.000]	[0.000]	[0.015]	[0.030]	[0.000]	[0.000]
Controls	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	3.18	3.18	0.20	0.20
Control SD	1.00	1.00	1.25	1.25	0.40	0.40
R^2	0.22	0.25	0.22	0.26	0.20	0.23
Observations	4541	4541	4541	4541	4541	4541

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 3a.

Table F2: Treatment knowledge

	ICW: Treatment knowledge		Fact-check quiz knowledge	
	(1)	(2)	(3)	(4)
A. Pooled estimation				
Placebo incentives	0.112	0.133	0.159	0.186
	(0.047)	(0.046)	(0.067)	(0.066)
	[0.009]	[0.002]	[0.009]	[0.002]
Pooled treatment	0.411	0.411	0.584	0.584
	(0.034)	(0.033)	(0.048)	(0.047)
	[0.000]	[0.000]	[0.000]	[0.000]
B. Disaggregated estimation				
Placebo incentives	0.113	0.132	0.160	0.187
	(0.047)	(0.046)	(0.067)	(0.066)
	[0.008]	[0.002]	[0.008]	[0.002]
Text information	0.335	0.345	0.476	0.489
	(0.064)	(0.061)	(0.091)	(0.087)
	[0.000]	[0.000]	[0.000]	[0.000]
Short podcast	0.388	0.379	0.551	0.538
	(0.046)	(0.045)	(0.065)	(0.064)
	[0.000]	[0.000]	[0.000]	[0.000]
Long podcast	0.373	0.386	0.529	0.548
	(0.048)	(0.046)	(0.068)	(0.065)
	[0.000]	[0.000]	[0.000]	[0.000]
Empathetic podcast	0.509	0.503	0.722	0.714
	(0.047)	(0.046)	(0.066)	(0.065)
	[0.000]	[0.000]	[0.000]	[0.000]
Controls	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓
Control Mean	0.00	0.00	2.40	2.40
Control SD	1.00	1.00	1.42	1.42
R^2	0.22	0.27	0.22	0.27
Observations	4541	4541	4541	4541

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 3b.

Table F3: Future take-up

	ICW: Future take-up		Stay subscribed to WCW		Want AC fact checks		Want AC reminders		Want AC vaccine info	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Pooled estimation										
Placebo incentives	0.061	0.058	0.013	0.011	-0.003	-0.002	0.030	0.029	0.049	0.047
	(0.050)	(0.048)	(0.021)	(0.021)	(0.019)	(0.019)	(0.023)	(0.023)	(0.023)	(0.023)
	[0.112]	[0.116]	[0.268]	[0.302]	[0.884]	[0.898]	[0.097]	[0.100]	[0.016]	[0.018]
Pooled treatment	0.205	0.207	0.140	0.139	0.052	0.053	0.082	0.083	0.092	0.092
	(0.034)	(0.033)	(0.014)	(0.014)	(0.013)	(0.013)	(0.016)	(0.016)	(0.016)	(0.016)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
B. Disaggregated estimation										
Placebo incentives	0.061	0.058	0.013	0.011	-0.003	-0.002	0.030	0.029	0.050	0.049
	(0.050)	(0.048)	(0.021)	(0.021)	(0.019)	(0.019)	(0.023)	(0.023)	(0.023)	(0.023)
	[0.111]	[0.116]	[0.265]	[0.305]	[0.885]	[0.900]	[0.096]	[0.100]	[0.016]	[0.015]
Text information	0.214	0.235	0.019	0.022	0.065	0.072	0.081	0.091	0.084	0.091
	(0.057)	(0.055)	(0.026)	(0.026)	(0.021)	(0.020)	(0.028)	(0.027)	(0.028)	(0.028)
	[0.000]	[0.000]	[0.230]	[0.195]	[0.001]	[0.000]	[0.002]	[0.000]	[0.001]	[0.001]
Short podcast	0.234	0.239	0.150	0.150	0.061	0.063	0.094	0.097	0.103	0.105
	(0.044)	(0.043)	(0.017)	(0.017)	(0.016)	(0.016)	(0.021)	(0.020)	(0.021)	(0.020)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Long podcast	0.172	0.171	0.168	0.166	0.039	0.040	0.069	0.068	0.085	0.085
	(0.045)	(0.044)	(0.016)	(0.016)	(0.017)	(0.016)	(0.021)	(0.021)	(0.021)	(0.021)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.009]	[0.008]	[0.001]	[0.001]	[0.000]	[0.000]
Empathetic podcast	0.202	0.196	0.156	0.153	0.049	0.048	0.083	0.080	0.093	0.090
	(0.044)	(0.043)	(0.017)	(0.017)	(0.017)	(0.016)	(0.021)	(0.021)	(0.021)	(0.021)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.002]	[0.002]	[0.000]	[0.000]	[0.000]	[0.000]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	0.75	0.75	0.82	0.82	0.66	0.66	0.66	0.66
Control SD	1.00	1.00	0.43	0.43	0.38	0.38	0.47	0.47	0.48	0.48
R^2	0.09	0.14	0.11	0.15	0.08	0.11	0.08	0.13	0.08	0.12
Observations	4541	4541	4541	4541	4541	4541	4541	4541	4541	4541

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 3c.

Table F4: Discernment

	ICW: Discernment		Alcohol and COVID (true)		Foreign restaurant workers (false)		How COVID spreads (true)		Matric marks (false)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Pooled estimation										
Placebo incentives	0.045	0.055	-0.020	-0.018	0.049	0.043	0.066	0.075	0.035	0.036
	(0.050)	(0.049)	(0.065)	(0.065)	(0.067)	(0.066)	(0.048)	(0.048)	(0.071)	(0.070)
	[0.180]	[0.130]	[0.758]	[0.776]	[0.234]	[0.255]	[0.085]	[0.060]	[0.311]	[0.303]
Pooled treatment	0.058	0.061	-0.126	-0.121	0.175	0.174	0.050	0.056	0.062	0.062
	(0.035)	(0.034)	(0.046)	(0.046)	(0.048)	(0.047)	(0.034)	(0.034)	(0.049)	(0.048)
	[0.048]	[0.039]	[0.007]	[0.009]	[0.000]	[0.000]	[0.072]	[0.050]	[0.102]	[0.098]
B. Disaggregated estimation										
Placebo incentives	0.046	0.056	-0.020	-0.014	0.049	0.043	0.066	0.076	0.035	0.036
	(0.050)	(0.049)	(0.065)	(0.065)	(0.067)	(0.066)	(0.048)	(0.048)	(0.071)	(0.070)
	[0.178]	[0.127]	[0.755]	[0.827]	[0.233]	[0.255]	[0.085]	[0.057]	[0.310]	[0.304]
Text information	0.120	0.120	-0.002	0.014	0.193	0.175	0.061	0.072	0.044	0.029
	(0.063)	(0.062)	(0.079)	(0.079)	(0.081)	(0.079)	(0.057)	(0.058)	(0.088)	(0.087)
	[0.029]	[0.026]	[0.982]	[0.432]	[0.009]	[0.013]	[0.146]	[0.106]	[0.309]	[0.369]
Short podcast	0.025	0.021	-0.155	-0.147	0.151	0.146	0.052	0.051	0.023	0.022
	(0.046)	(0.045)	(0.061)	(0.061)	(0.062)	(0.060)	(0.043)	(0.043)	(0.062)	(0.061)
	[0.289]	[0.317]	[0.012]	[0.017]	[0.007]	[0.008]	[0.112]	[0.114]	[0.359]	[0.360]
Long podcast	-0.018	-0.003	-0.161	-0.153	0.085	0.092	0.047	0.057	-0.020	-0.012
	(0.046)	(0.046)	(0.063)	(0.063)	(0.064)	(0.063)	(0.046)	(0.046)	(0.066)	(0.065)
	[0.691]	[0.945]	[0.010]	[0.015]	[0.092]	[0.073]	[0.151]	[0.106]	[0.767]	[0.854]
Empathetic podcast	0.141	0.143	-0.119	-0.109	0.280	0.287	0.045	0.053	0.194	0.194
	(0.046)	(0.045)	(0.061)	(0.061)	(0.063)	(0.062)	(0.045)	(0.045)	(0.063)	(0.063)
	[0.001]	[0.001]	[0.053]	[0.074]	[0.000]	[0.000]	[0.158]	[0.120]	[0.001]	[0.001]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	-2.41	-2.41	2.78	2.78	-1.58	-1.58	3.07	3.07
Control SD	1.00	1.00	1.27	1.27	1.32	1.32	0.97	0.97	1.35	1.35
R^2	0.08	0.13	0.08	0.09	0.11	0.16	0.08	0.10	0.10	0.14
Observations	4541	4541	4143	4143	4143	4143	4143	4143	4143	4143

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 4a.

Table F5: Identification of conspiracy theories

	ICW: Conspiracy theories		AIDS intentionally created (reversed)		Nelson Mandela died in 1985 (reversed)		Vaccines cause infertility (reversed)		Vaccines have microchips (reversed)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Pooled estimation										
Placebo incentives	-0.024	-0.003	-0.095	-0.078	0.013	0.028	0.015	0.039	-0.012	0.009
	(0.050)	(0.048)	(0.070)	(0.068)	(0.070)	(0.068)	(0.067)	(0.066)	(0.069)	(0.068)
	[0.635]	[0.947]	[0.170]	[0.255]	[0.427]	[0.339]	[0.413]	[0.276]	[0.867]	[0.450]
Pooled treatment	0.104	0.109	0.071	0.079	0.093	0.098	0.177	0.183	0.110	0.110
	(0.035)	(0.034)	(0.048)	(0.048)	(0.048)	(0.047)	(0.048)	(0.047)	(0.047)	(0.047)
	[0.001]	[0.001]	[0.071]	[0.048]	[0.026]	[0.018]	[0.000]	[0.000]	[0.010]	[0.009]
B. Disaggregated estimation										
Placebo incentives	-0.024	-0.003	-0.095	-0.079	0.013	0.029	0.015	0.041	-0.012	0.009
	(0.050)	(0.049)	(0.070)	(0.068)	(0.070)	(0.068)	(0.068)	(0.066)	(0.069)	(0.068)
	[0.637]	[0.954]	[0.170]	[0.248]	[0.426]	[0.336]	[0.411]	[0.269]	[0.868]	[0.448]
Text information	0.106	0.110	0.103	0.108	0.085	0.088	0.132	0.132	0.133	0.134
	(0.058)	(0.058)	(0.085)	(0.084)	(0.080)	(0.079)	(0.082)	(0.082)	(0.078)	(0.079)
	[0.034]	[0.029]	[0.113]	[0.100]	[0.143]	[0.134]	[0.053]	[0.054]	[0.045]	[0.045]
Short podcast	0.039	0.039	0.000	-0.004	0.064	0.066	0.061	0.065	0.052	0.050
	(0.046)	(0.045)	(0.064)	(0.063)	(0.064)	(0.062)	(0.063)	(0.062)	(0.062)	(0.061)
	[0.199]	[0.189]	[1.000]	[0.947]	[0.157]	[0.145]	[0.166]	[0.145]	[0.202]	[0.209]
Long podcast	0.109	0.126	0.082	0.104	0.089	0.111	0.190	0.206	0.108	0.120
	(0.046)	(0.044)	(0.064)	(0.062)	(0.064)	(0.062)	(0.063)	(0.061)	(0.063)	(0.062)
	[0.009]	[0.002]	[0.100]	[0.047]	[0.083]	[0.036]	[0.001]	[0.000]	[0.043]	[0.026]
Empathetic podcast	0.166	0.163	0.119	0.126	0.132	0.124	0.306	0.301	0.163	0.152
	(0.045)	(0.043)	(0.063)	(0.062)	(0.063)	(0.062)	(0.060)	(0.059)	(0.061)	(0.060)
	[0.000]	[0.000]	[0.029]	[0.021]	[0.018]	[0.022]	[0.000]	[0.000]	[0.004]	[0.006]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	-2.34	-2.34	-2.24	-2.24	-2.39	-2.39	-2.36	-2.36
Control SD	1.00	1.00	1.38	1.38	1.36	1.36	1.35	1.35	1.35	1.35
R^2	0.09	0.16	0.08	0.12	0.08	0.15	0.08	0.12	0.07	0.12
Observations	4541	4541	4541	4541	4541	4541	4541	4541	4541	4541

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while *p*-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 4b.

Table F6: Knowledge of verification methods (part 1)

	ICW: Verification knowledge		Avoid misinfo: Ask others (reversed)		Avoid misinfo: Seek reputable orgs		How verify (use sources)		Strategy: Ask experts		Strategy: Ask themselves (reversed)		Strategy: Check popular source (reversed)		Strategy: Talk to others (reversed)		Strategy: Use image search	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
A. Pooled estimation																		
Placebo incentives	0.039	0.048	0.012	0.010	0.025	0.028	-0.028	-0.028	0.022	0.025	-0.021	-0.021	-0.011	-0.009	0.002	0.004	0.035	0.034
	(0.050)	(0.050)	(0.018)	(0.018)	(0.024)	(0.023)	(0.051)	(0.049)	(0.025)	(0.024)	(0.017)	(0.017)	(0.025)	(0.024)	(0.019)	(0.019)	(0.017)	(0.017)
	[0.216]	[0.167]	[0.245]	[0.287]	[0.149]	[0.114]	[0.579]	[0.571]	[0.185]	[0.153]	[0.198]	[0.200]	[0.654]	[0.699]	[0.460]	[0.417]	[0.022]	[0.025]
Pooled treatment	0.096	0.099	-0.020	-0.017	0.031	0.034	0.026	0.030	0.049	0.050	-0.013	-0.015	-0.014	-0.016	-0.001	-0.003	0.070	0.070
	(0.036)	(0.036)	(0.012)	(0.012)	(0.017)	(0.017)	(0.036)	(0.035)	(0.017)	(0.017)	(0.011)	(0.011)	(0.017)	(0.017)	(0.014)	(0.013)	(0.012)	(0.011)
	[0.003]	[0.003]	[0.099]	[0.142]	[0.031]	[0.022]	[0.231]	[0.195]	[0.002]	[0.002]	[0.235]	[0.161]	[0.396]	[0.331]	[0.955]	[0.824]	[0.000]	[0.000]
B. Disaggregated estimation																		
Placebo incentives	0.039	0.048	0.012	0.010	0.025	0.027	-0.027	-0.027	0.022	0.024	-0.021	-0.022	-0.011	-0.009	0.002	0.003	0.035	0.034
	(0.050)	(0.050)	(0.018)	(0.018)	(0.024)	(0.023)	(0.051)	(0.049)	(0.025)	(0.024)	(0.017)	(0.017)	(0.025)	(0.024)	(0.019)	(0.019)	(0.017)	(0.017)
	[0.214]	[0.166]	[0.246]	[0.282]	[0.149]	[0.125]	[0.588]	[0.576]	[0.186]	[0.159]	[0.199]	[0.187]	[0.654]	[0.703]	[0.457]	[0.430]	[0.022]	[0.025]
Text information	0.167	0.173	0.011	0.012	0.036	0.038	-0.031	-0.027	0.071	0.071	-0.031	-0.033	-0.009	-0.013	0.011	0.010	0.039	0.038
	(0.064)	(0.064)	(0.022)	(0.022)	(0.030)	(0.030)	(0.064)	(0.060)	(0.031)	(0.030)	(0.020)	(0.020)	(0.030)	(0.030)	(0.024)	(0.023)	(0.021)	(0.021)
	[0.005]	[0.004]	[0.316]	[0.295]	[0.120]	[0.100]	[0.625]	[0.649]	[0.010]	[0.009]	[0.132]	[0.103]	[0.762]	[0.651]	[0.325]	[0.339]	[0.033]	[0.033]
Short podcast	0.124	0.118	-0.003	-0.002	0.010	0.008	0.061	0.054	0.034	0.032	0.001	0.001	-0.006	-0.006	-0.010	-0.010	0.076	0.074
	(0.048)	(0.048)	(0.016)	(0.016)	(0.022)	(0.022)	(0.047)	(0.046)	(0.023)	(0.022)	(0.014)	(0.014)	(0.022)	(0.022)	(0.018)	(0.018)	(0.016)	(0.016)
	[0.005]	[0.007]	[0.872]	[0.893]	[0.320]	[0.351]	[0.098]	[0.117]	[0.069]	[0.075]	[0.469]	[0.916]	[0.804]	[0.790]	[0.597]	[0.594]	[0.000]	[0.000]
Long podcast	0.022	0.034	-0.032	-0.030	0.035	0.040	-0.030	-0.016	0.056	0.061	-0.012	-0.013	-0.018	-0.021	-0.018	-0.022	0.063	0.065
	(0.048)	(0.048)	(0.015)	(0.015)	(0.023)	(0.022)	(0.047)	(0.046)	(0.023)	(0.022)	(0.015)	(0.015)	(0.023)	(0.023)	(0.018)	(0.018)	(0.016)	(0.016)
	[0.324]	[0.240]	[0.034]	[0.049]	[0.062]	[0.035]	[0.533]	[0.728]	[0.007]	[0.003]	[0.415]	[0.363]	[0.438]	[0.353]	[0.325]	[0.226]	[0.000]	[0.000]
Empathetic podcast	0.110	0.109	-0.039	-0.033	0.048	0.050	0.073	0.076	0.046	0.046	-0.021	-0.021	-0.023	-0.024	0.020	0.016	0.087	0.086
	(0.049)	(0.049)	(0.015)	(0.015)	(0.023)	(0.022)	(0.048)	(0.047)	(0.023)	(0.023)	(0.015)	(0.015)	(0.022)	(0.022)	(0.018)	(0.017)	(0.017)	(0.017)
	[0.012]	[0.013]	[0.010]	[0.025]	[0.017]	[0.013]	[0.065]	[0.052]	[0.021]	[0.020]	[0.164]	[0.152]	[0.311]	[0.275]	[0.122]	[0.172]	[0.000]	[0.000]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	0.14	0.14	0.34	0.34	0.00	0.00	0.39	0.39	-0.11	-0.11	-0.36	-0.36	-0.18	-0.18	0.11	0.11
Control SD	1.00	1.00	0.35	0.35	0.48	0.48	1.00	1.00	0.49	0.49	0.31	0.31	0.48	0.48	0.38	0.38	0.31	0.31
R^2	0.09	0.11	0.06	0.09	0.06	0.09	0.07	0.15	0.07	0.12	0.07	0.10	0.06	0.08	0.06	0.09	0.09	0.14
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 5a.

Table F7: Knowledge of verification methods (part 2)

	ICW: Verification knowledge		To verify: Ask family on WA (reversed)		To verify: Ask in person (reversed)		To verify: Ask others on WA (reversed)		To verify: Post on social media (reversed)		To verify: Submit fact-check request		To verify: Use fact-checker		To verify: Use internet	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
A. Pooled estimation																
Placebo incentives	0.039	0.046	0.006	0.008	0.000	0.003	0.008	0.011	-0.017	-0.013	0.017	0.020	0.026	0.025	-0.023	-0.019
	(0.050)	(0.050)	(0.019)	(0.018)	(0.023)	(0.023)	(0.015)	(0.015)	(0.017)	(0.017)	(0.020)	(0.019)	(0.025)	(0.024)	(0.024)	(0.024)
	[0.216]	[0.177]	[0.370]	[0.334]	[0.991]	[0.452]	[0.283]	[0.222]	[0.321]	[0.431]	[0.190]	[0.150]	[0.144]	[0.150]	[0.338]	[0.417]
Pooled treatment	0.096	0.098	-0.014	-0.013	0.027	0.025	0.005	0.005	0.007	0.006	0.050	0.049	0.053	0.055	-0.012	-0.007
	(0.036)	(0.036)	(0.013)	(0.013)	(0.016)	(0.016)	(0.010)	(0.010)	(0.012)	(0.011)	(0.014)	(0.014)	(0.017)	(0.017)	(0.017)	(0.017)
	[0.003]	[0.003]	[0.315]	[0.316]	[0.045]	[0.055]	[0.311]	[0.317]	[0.275]	[0.300]	[0.000]	[0.000]	[0.001]	[0.001]	[0.495]	[0.681]
B. Disaggregated estimation																
Placebo incentives	0.039	0.046	0.006	0.008	0.000	0.003	0.008	0.011	-0.017	-0.014	0.017	0.020	0.026	0.026	-0.023	-0.020
	(0.050)	(0.050)	(0.019)	(0.018)	(0.023)	(0.023)	(0.015)	(0.015)	(0.017)	(0.017)	(0.020)	(0.019)	(0.025)	(0.024)	(0.024)	(0.024)
	[0.214]	[0.176]	[0.370]	[0.339]	[0.991]	[0.451]	[0.281]	[0.223]	[0.319]	[0.405]	[0.189]	[0.155]	[0.144]	[0.142]	[0.342]	[0.415]
Text information	0.167	0.174	-0.007	-0.007	0.056	0.055	0.010	0.009	0.026	0.028	0.075	0.074	0.087	0.089	-0.016	-0.008
	(0.064)	(0.064)	(0.024)	(0.023)	(0.027)	(0.026)	(0.018)	(0.017)	(0.019)	(0.019)	(0.026)	(0.026)	(0.030)	(0.030)	(0.030)	(0.030)
	[0.005]	[0.003]	[0.765]	[0.748]	[0.020]	[0.019]	[0.294]	[0.308]	[0.085]	[0.071]	[0.002]	[0.002]	[0.002]	[0.001]	[0.592]	[0.780]
Short podcast	0.124	0.119	-0.011	-0.011	0.002	0.001	0.004	0.003	0.010	0.010	0.059	0.056	0.053	0.050	0.001	0.003
	(0.048)	(0.048)	(0.018)	(0.018)	(0.021)	(0.021)	(0.013)	(0.013)	(0.015)	(0.015)	(0.019)	(0.019)	(0.023)	(0.023)	(0.023)	(0.022)
	[0.005]	[0.006]	[0.552]	[0.536]	[0.469]	[0.474]	[0.397]	[0.423]	[0.244]	[0.256]	[0.001]	[0.001]	[0.011]	[0.014]	[0.487]	[0.446]
Long podcast	0.022	0.033	-0.019	-0.020	0.017	0.014	-0.008	-0.006	0.007	0.005	0.027	0.028	0.046	0.052	-0.034	-0.026
	(0.048)	(0.048)	(0.018)	(0.018)	(0.021)	(0.021)	(0.014)	(0.014)	(0.015)	(0.015)	(0.018)	(0.018)	(0.023)	(0.023)	(0.023)	(0.022)
	[0.324]	[0.245]	[0.293]	[0.267]	[0.206]	[0.250]	[0.560]	[0.663]	[0.327]	[0.367]	[0.071]	[0.058]	[0.025]	[0.012]	[0.133]	[0.251]
Empathetic podcast	0.110	0.109	-0.014	-0.013	0.050	0.048	0.018	0.017	-0.005	-0.007	0.052	0.050	0.046	0.050	0.000	0.000
	(0.049)	(0.049)	(0.018)	(0.017)	(0.021)	(0.020)	(0.013)	(0.013)	(0.015)	(0.015)	(0.019)	(0.019)	(0.024)	(0.023)	(0.023)	(0.023)
	[0.012]	[0.014]	[0.432]	[0.467]	[0.007]	[0.009]	[0.087]	[0.095]	[0.724]	[0.635]	[0.003]	[0.004]	[0.024]	[0.016]	[1.000]	[0.496]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	-0.18	-0.18	-0.31	-0.31	-0.1	-0.1	-0.13	-0.13	0.18	0.18	0.46	0.46	0.47	0.47
Control SD	1.00	1.00	0.38	0.38	0.46	0.46	0.30	0.30	0.33	0.33	0.38	0.38	0.50	0.50	0.50	0.50
R^2	0.09	0.10	0.08	0.11	0.09	0.14	0.08	0.10	0.07	0.09	0.07	0.11	0.08	0.11	0.11	0.14
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 5a.

Table F8: Trust in social media (besides WhatsApp)

	ICW: Trust social media		How true: Info from other social media		Trust most for info: Other social media		Trust: Info from other social media	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Pooled estimation								
Placebo incentives	-0.035	-0.045	0.004	-0.005	-0.023	-0.023	-0.014	-0.027
	(0.047)	(0.047)	(0.038)	(0.036)	(0.019)	(0.018)	(0.050)	(0.049)
	[0.226]	[0.168]	[0.910]	[0.450]	[0.111]	[0.101]	[0.387]	[0.294]
Pooled treatment	-0.088	-0.086	-0.049	-0.043	-0.035	-0.031	-0.049	-0.050
	(0.034)	(0.033)	(0.026)	(0.025)	(0.014)	(0.013)	(0.035)	(0.035)
	[0.004]	[0.004]	[0.028]	[0.043]	[0.005]	[0.009]	[0.083]	[0.073]
B. Disaggregated estimation								
Placebo incentives	-0.036	-0.046	0.004	-0.005	-0.023	-0.023	-0.015	-0.027
	(0.047)	(0.047)	(0.038)	(0.036)	(0.019)	(0.018)	(0.050)	(0.050)
	[0.226]	[0.163]	[0.912]	[0.446]	[0.111]	[0.111]	[0.385]	[0.290]
Text information	-0.153	-0.138	-0.102	-0.085	-0.055	-0.049	-0.066	-0.054
	(0.058)	(0.056)	(0.044)	(0.043)	(0.022)	(0.022)	(0.062)	(0.061)
	[0.004]	[0.007]	[0.011]	[0.023]	[0.007]	[0.012]	[0.144]	[0.185]
Short podcast	-0.023	-0.024	-0.024	-0.015	-0.010	-0.006	-0.007	-0.015
	(0.044)	(0.043)	(0.034)	(0.032)	(0.018)	(0.018)	(0.046)	(0.045)
	[0.303]	[0.289]	[0.234]	[0.318]	[0.278]	[0.369]	[0.439]	[0.367]
Long podcast	-0.067	-0.071	-0.023	-0.027	-0.033	-0.031	-0.030	-0.039
	(0.045)	(0.044)	(0.035)	(0.034)	(0.018)	(0.017)	(0.047)	(0.047)
	[0.065]	[0.052]	[0.253]	[0.212]	[0.032]	[0.038]	[0.262]	[0.199]
Empathetic podcast	-0.148	-0.142	-0.076	-0.068	-0.052	-0.048	-0.103	-0.099
	(0.043)	(0.043)	(0.034)	(0.032)	(0.017)	(0.017)	(0.046)	(0.045)
	[0.000]	[0.000]	[0.012]	[0.018]	[0.001]	[0.002]	[0.013]	[0.014]
Controls	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	2.87	2.87	0.19	0.19	2.91	2.91
Control SD	1.00	1.00	0.73	0.73	0.39	0.39	1.04	1.04
R^2	0.14	0.18	0.10	0.18	0.07	0.10	0.14	0.17
Observations	4541	4541	4541	4541	4541	4541	4541	4541

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 5b.

Table F9: Social media consumption

	ICW: Consume social media		Get news from: Other social media	
	(1)	(2)	(3)	(4)
A. Pooled estimation				
Placebo incentives	-0.015	-0.022	-0.015	-0.015
	(0.049)	(0.048)	(0.024)	(0.024)
	[0.381]	[0.326]	[0.265]	[0.270]
Pooled treatment	-0.004	-0.007	-0.008	-0.007
	(0.034)	(0.034)	(0.017)	(0.017)
	[0.453]	[0.416]	[0.323]	[0.335]
B. Disaggregated estimation				
Placebo incentives	-0.015	-0.022	-0.015	-0.015
	(0.049)	(0.048)	(0.024)	(0.024)
	[0.381]	[0.327]	[0.266]	[0.271]
Text information	-0.071	-0.069	-0.037	-0.036
	(0.060)	(0.060)	(0.030)	(0.030)
	[0.120]	[0.123]	[0.107]	[0.112]
Short podcast	0.022	0.024	0.008	0.010
	(0.045)	(0.045)	(0.023)	(0.022)
	[0.622]	[0.599]	[0.732]	[0.663]
Long podcast	0.023	0.013	0.002	0.000
	(0.045)	(0.045)	(0.023)	(0.022)
	[0.607]	[0.767]	[0.940]	[0.989]
Empathetic podcast	-0.028	-0.031	-0.020	-0.019
	(0.045)	(0.044)	(0.022)	(0.022)
	[0.263]	[0.240]	[0.185]	[0.195]
Controls	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓
Control Mean	0.00	0.00	0.43	0.43
Control SD	1.00	1.00	0.50	0.50
R^2	0.12	0.14	0.10	0.13
Observations	4541	4541	4541	4541

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 6a.

Table F10: Active verification

	ICW: Active verification		How often verify	
	(1)	(2)	(3)	(4)
A. Pooled estimation				
Placebo incentives	-0.039	-0.038	-0.043	-0.042
	(0.048)	(0.048)	(0.054)	(0.053)
	[0.419]	[0.435]	[0.419]	[0.435]
Pooled treatment	-0.038	-0.039	-0.042	-0.043
	(0.034)	(0.034)	(0.038)	(0.037)
	[0.271]	[0.252]	[0.271]	[0.252]
B. Disaggregated estimation				
Placebo incentives	-0.039	-0.040	-0.044	-0.042
	(0.048)	(0.048)	(0.054)	(0.053)
	[0.417]	[0.403]	[0.417]	[0.434]
Text information	-0.127	-0.126	-0.141	-0.141
	(0.065)	(0.064)	(0.072)	(0.071)
	[0.050]	[0.048]	[0.050]	[0.046]
Short podcast	-0.042	-0.043	-0.046	-0.047
	(0.045)	(0.044)	(0.049)	(0.049)
	[0.351]	[0.334]	[0.351]	[0.336]
Long podcast	0.016	0.015	0.018	0.015
	(0.043)	(0.043)	(0.048)	(0.048)
	[0.357]	[0.364]	[0.357]	[0.375]
Empathetic podcast	-0.046	-0.047	-0.051	-0.052
	(0.046)	(0.045)	(0.051)	(0.050)
	[0.312]	[0.303]	[0.312]	[0.300]
Controls	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓
Control Mean	0.00	0.00	3.86	3.86
Control SD	1.00	1.00	1.11	1.11
R^2	0.11	0.14	0.11	0.14
Observations	4541	4541	4541	4541

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 6b.

Table F11: Sharing

	ICW: Sharing		How often share stories	
	(1)	(2)	(3)	(4)
A. Pooled estimation				
Placebo incentives	0.022	0.004	0.023	0.001
	(0.046)	(0.045)	(0.054)	(0.051)
	[0.630]	[0.928]	[0.673]	[0.495]
Pooled treatment	-0.027	-0.029	-0.033	-0.033
	(0.033)	(0.032)	(0.039)	(0.037)
	[0.206]	[0.184]	[0.194]	[0.182]
B. Disaggregated estimation				
Placebo incentives	0.022	0.004	0.023	-0.001
	(0.046)	(0.045)	(0.054)	(0.051)
	[0.630]	[0.932]	[0.675]	[0.991]
Text information	-0.101	-0.093	-0.118	-0.104
	(0.057)	(0.054)	(0.065)	(0.062)
	[0.038]	[0.044]	[0.034]	[0.046]
Short podcast	0.022	0.017	0.025	0.021
	(0.044)	(0.042)	(0.051)	(0.049)
	[0.613]	[0.687]	[0.628]	[0.658]
Long podcast	-0.001	-0.010	0.006	-0.009
	(0.044)	(0.043)	(0.051)	(0.049)
	[0.487]	[0.410]	[0.900]	[0.429]
Empathetic podcast	-0.070	-0.068	-0.095	-0.085
	(0.043)	(0.041)	(0.050)	(0.048)
	[0.050]	[0.050]	[0.029]	[0.037]
Controls	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓
Control Mean	0.00	0.00	2.85	2.85
Control SD	1.00	1.00	1.13	1.13
R^2	0.17	0.24	0.12	0.22
Observations	4541	4541	4541	4541

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 6c.

Table F12: COVID-19 beliefs and preventative behavior

	ICW: COVID-19 beliefs and behavior		Behavior: Stayed home		Behavior: Visited indoors (reversed)		Behavior: Wore mask		COVID is a hoax (reversed)		Lockdowns unnecessary (reversed)		Trust vaccines		Would get vaccinated	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
A. Pooled estimation																
Placebo incentives	-0.041	-0.037	-0.068	-0.075	-0.108	-0.099	0.169	0.175	0.068	0.081	-0.041	-0.028	-0.043	-0.029	-0.031	-0.028
	(0.048)	(0.048)	(0.107)	(0.106)	(0.103)	(0.101)	(0.114)	(0.114)	(0.055)	(0.054)	(0.045)	(0.045)	(0.068)	(0.067)	(0.078)	(0.077)
	[0.389]	[0.443]	[0.527]	[0.479]	[0.295]	[0.327]	[0.070]	[0.062]	[0.109]	[0.068]	[0.367]	[0.529]	[0.531]	[0.671]	[0.691]	[0.715]
Pooled treatment	0.003	0.006	-0.030	-0.026	-0.027	-0.023	0.049	0.054	0.084	0.091	-0.017	-0.009	0.025	0.030	0.041	0.045
	(0.034)	(0.033)	(0.076)	(0.076)	(0.071)	(0.070)	(0.080)	(0.080)	(0.040)	(0.039)	(0.032)	(0.032)	(0.049)	(0.048)	(0.055)	(0.054)
	[0.469]	[0.432]	[0.696]	[0.728]	[0.703]	[0.745]	[0.273]	[0.251]	[0.017]	[0.010]	[0.594]	[0.788]	[0.304]	[0.267]	[0.231]	[0.206]
B. Disaggregated estimation																
Placebo incentives	-0.042	-0.035	-0.068	-0.078	-0.108	-0.100	0.167	0.174	0.068	0.080	-0.040	-0.029	-0.043	-0.029	-0.032	-0.029
	(0.048)	(0.048)	(0.107)	(0.107)	(0.103)	(0.101)	(0.114)	(0.114)	(0.055)	(0.054)	(0.045)	(0.045)	(0.068)	(0.067)	(0.078)	(0.077)
	[0.384]	[0.463]	[0.524]	[0.464]	[0.294]	[0.323]	[0.071]	[0.063]	[0.109]	[0.072]	[0.372]	[0.523]	[0.528]	[0.668]	[0.684]	[0.708]
Text information	0.142	0.153	0.054	0.052	0.265	0.275	0.271	0.295	0.093	0.096	-0.063	-0.049	0.048	0.073	0.121	0.142
	(0.057)	(0.057)	(0.131)	(0.130)	(0.124)	(0.122)	(0.129)	(0.128)	(0.067)	(0.067)	(0.057)	(0.056)	(0.084)	(0.082)	(0.093)	(0.092)
	[0.007]	[0.004]	[0.341]	[0.345]	[0.016]	[0.012]	[0.018]	[0.011]	[0.084]	[0.076]	[0.266]	[0.383]	[0.284]	[0.186]	[0.096]	[0.062]
Short podcast	0.019	0.022	-0.003	0.002	-0.033	-0.027	0.090	0.087	0.114	0.116	0.040	0.045	0.054	0.054	0.053	0.054
	(0.044)	(0.043)	(0.101)	(0.101)	(0.094)	(0.093)	(0.105)	(0.104)	(0.051)	(0.050)	(0.042)	(0.041)	(0.064)	(0.063)	(0.072)	(0.072)
	[0.330]	[0.303]	[0.973]	[0.494]	[0.726]	[0.767]	[0.195]	[0.201]	[0.012]	[0.010]	[0.167]	[0.135]	[0.198]	[0.195]	[0.230]	[0.225]
Long podcast	-0.025	-0.018	-0.016	-0.019	-0.126	-0.111	0.067	0.073	0.060	0.074	-0.057	-0.046	0.044	0.050	0.089	0.090
	(0.047)	(0.046)	(0.101)	(0.101)	(0.099)	(0.097)	(0.106)	(0.106)	(0.052)	(0.052)	(0.043)	(0.043)	(0.065)	(0.064)	(0.072)	(0.071)
	[0.599]	[0.694]	[0.875]	[0.848]	[0.201]	[0.253]	[0.264]	[0.245]	[0.125]	[0.076]	[0.186]	[0.284]	[0.252]	[0.219]	[0.109]	[0.103]
Empathetic podcast	-0.051	-0.056	-0.108	-0.102	-0.055	-0.064	-0.116	-0.111	0.072	0.072	-0.015	-0.006	-0.034	-0.036	-0.058	-0.056
	(0.045)	(0.044)	(0.101)	(0.101)	(0.095)	(0.094)	(0.109)	(0.108)	(0.052)	(0.051)	(0.042)	(0.042)	(0.064)	(0.063)	(0.073)	(0.072)
	[0.253]	[0.206]	[0.282]	[0.313]	[0.562]	[0.494]	[0.288]	[0.307]	[0.082]	[0.079]	[0.720]	[0.882]	[0.594]	[0.572]	[0.427]	[0.441]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	4.25	4.25	-1.75	-1.75	5.23	5.23	-1.7	-1.7	-1.77	-1.77	3.37	3.37	3.46	3.46
Control SD	1.00	1.00	2.25	2.25	2.05	2.05	2.41	2.41	1.14	1.14	0.92	0.92	1.39	1.39	1.57	1.57
R^2	0.11	0.15	0.15	0.16	0.10	0.13	0.14	0.15	0.08	0.11	0.09	0.11	0.07	0.11	0.06	0.09
Observations	4541	4541	4541	4541	4541	4541	4541	4541	4541	4541	4541	4541	4541	4541	4541	4541

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 7a.

Table F13: Views and attitudes about the government

	ICW: Government attitudes		General govt performance		Govt handled COVID well		How true: Info from politicians		Trust most for info: Govt		Trust most for info: Politicians		Trust: Info from politicians		Vote: Local incumbent	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
A. Pooled estimation																
Placebo incentives	0.097	0.090	0.079	0.074	0.026	0.022	-0.030	-0.034	0.009	0.009	0.019	0.018	-0.013	-0.024	0.060	0.057
	(0.050)	(0.047)	(0.059)	(0.056)	(0.061)	(0.060)	(0.048)	(0.046)	(0.023)	(0.022)	(0.017)	(0.017)	(0.059)	(0.056)	(0.021)	(0.021)
	[0.027]	[0.028]	[0.089]	[0.096]	[0.334]	[0.353]	[0.525]	[0.466]	[0.343]	[0.339]	[0.132]	[0.139]	[0.827]	[0.671]	[0.002]	[0.003]
Pooled treatment	0.061	0.058	0.051	0.042	0.027	0.029	-0.033	-0.032	0.021	0.021	0.020	0.020	-0.035	-0.035	0.020	0.020
	(0.035)	(0.033)	(0.042)	(0.040)	(0.043)	(0.042)	(0.033)	(0.032)	(0.016)	(0.016)	(0.012)	(0.011)	(0.041)	(0.040)	(0.014)	(0.014)
	[0.039]	[0.041]	[0.109]	[0.150]	[0.264]	[0.244]	[0.323]	[0.326]	[0.090]	[0.096]	[0.046]	[0.040]	[0.396]	[0.378]	[0.081]	[0.081]
B. Disaggregated estimation																
Placebo incentives	0.097	0.090	0.079	0.072	0.027	0.024	-0.030	-0.034	0.009	0.009	0.019	0.018	-0.013	-0.024	0.059	0.056
	(0.050)	(0.047)	(0.059)	(0.057)	(0.061)	(0.060)	(0.048)	(0.046)	(0.023)	(0.022)	(0.017)	(0.017)	(0.059)	(0.056)	(0.021)	(0.021)
	[0.027]	[0.028]	[0.089]	[0.103]	[0.331]	[0.344]	[0.528]	[0.459]	[0.342]	[0.341]	[0.131]	[0.139]	[0.827]	[0.669]	[0.003]	[0.004]
Text information	0.075	0.083	0.033	0.032	-0.062	-0.048	0.037	0.050	0.047	0.049	0.000	0.004	-0.009	0.006	0.055	0.059
	(0.061)	(0.058)	(0.069)	(0.068)	(0.074)	(0.072)	(0.057)	(0.056)	(0.030)	(0.029)	(0.020)	(0.020)	(0.076)	(0.074)	(0.026)	(0.026)
	[0.111]	[0.075]	[0.316]	[0.319]	[0.405]	[0.509]	[0.258]	[0.187]	[0.057]	[0.046]	[0.492]	[0.424]	[0.910]	[0.468]	[0.017]	[0.011]
Short podcast	0.121	0.114	0.095	0.085	0.118	0.112	0.015	0.016	0.032	0.028	0.026	0.027	0.017	0.013	0.032	0.029
	(0.046)	(0.044)	(0.055)	(0.053)	(0.056)	(0.055)	(0.043)	(0.042)	(0.021)	(0.021)	(0.015)	(0.015)	(0.054)	(0.052)	(0.019)	(0.019)
	[0.004]	[0.005]	[0.043]	[0.055]	[0.017]	[0.021]	[0.361]	[0.349]	[0.067]	[0.093]	[0.046]	[0.039]	[0.377]	[0.399]	[0.050]	[0.064]
Long podcast	0.040	0.030	0.032	0.014	-0.013	-0.013	-0.098	-0.108	0.000	0.001	0.020	0.019	-0.062	-0.073	0.034	0.032
	(0.048)	(0.046)	(0.056)	(0.055)	(0.057)	(0.056)	(0.045)	(0.044)	(0.021)	(0.021)	(0.016)	(0.016)	(0.056)	(0.055)	(0.020)	(0.019)
	[0.203]	[0.257]	[0.283]	[0.397]	[0.822]	[0.822]	[0.028]	[0.014]	[0.993]	[0.484]	[0.103]	[0.117]	[0.270]	[0.183]	[0.041]	[0.049]
Empathetic podcast	0.015	0.017	0.034	0.031	0.013	0.019	-0.049	-0.044	0.020	0.020	0.021	0.022	-0.075	-0.066	-0.023	-0.020
	(0.047)	(0.045)	(0.055)	(0.053)	(0.056)	(0.055)	(0.045)	(0.044)	(0.021)	(0.021)	(0.016)	(0.015)	(0.055)	(0.053)	(0.018)	(0.018)
	[0.376]	[0.354]	[0.269]	[0.278]	[0.407]	[0.366]	[0.276]	[0.316]	[0.169]	[0.173]	[0.085]	[0.078]	[0.173]	[0.213]	[0.219]	[0.276]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	2.33	2.33	3.07	3.07	3.04	3.04	0.29	0.29	0.12	0.12	2.91	2.91	0.21	0.21
Control SD	1.00	1.00	1.21	1.21	1.24	1.24	0.94	0.94	0.45	0.45	0.32	0.32	1.20	1.20	0.40	0.40
R^2	0.10	0.20	0.11	0.17	0.08	0.13	0.08	0.13	0.07	0.11	0.07	0.09	0.09	0.18	0.08	0.12
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table C1 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets. ICW estimate plotted in Figure 7b.

Table F14: Pooled estimation ICW outcomes (including LASSO-selected controls)

Figure	3a	3b	3c	4a	4b	5a	5b	6a	6b	6c	7a	7b
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Placebo incentives	0.32 (0.05)	0.13 (0.05)	0.06 (0.05)	0.06 (0.05)	0.00 (0.05)	0.05 (0.05)	-0.05 (0.05)	-0.02 (0.05)	-0.04 (0.05)	0.00 (0.04)	-0.03 (0.05)	0.09 (0.05)
Pooled treatment	0.56 (0.03)	0.41 (0.03)	0.21 (0.03)	0.06 (0.03)	0.11 (0.03)	0.10 (0.04)	-0.09 (0.03)	-0.01 (0.03)	-0.04 (0.03)	-0.03 (0.03)	0.01 (0.03)	0.06 (0.03)
Gender: Female	-0.06 (0.04)	-0.03 (0.04)	-0.01 (0.03)	-0.03 (0.04)		-0.01 (0.04)		0.01 (0.03)			0.11 (0.04)	-0.02 (0.04)
Locality: Peri-urban	0.02 (0.03)	0.04 (0.02)		0.08 (0.03)			-0.01 (0.02)	0.01 (0.03)				
Locality: Rural	0.07 (0.03)		0.00 (0.02)	0.03 (0.03)		-0.01 (0.02)		-0.02 (0.03)		-0.02 (0.02)		0.06 (0.02)
Age: 18-24	0.06 (0.02)	0.00 (0.04)	-0.08 (0.02)	-0.06 (0.04)	-0.21 (0.02)	0.03 (0.02)		-0.05 (0.02)	-0.05 (0.02)	-0.02 (0.02)	-0.16 (0.04)	0.18 (0.02)
Age: 25-34	0.09 (0.02)	0.01 (0.04)		-0.06 (0.04)	-0.11 (0.02)	0.01 (0.02)					-0.10 (0.04)	0.12 (0.02)
Education: Secondary	0.02 (0.02)	0.05 (0.01)	-0.01 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.02)	-0.02 (0.01)			-0.02 (0.01)		-0.02 (0.01)
Education: University	0.03 (0.02)	0.06 (0.02)	-0.06 (0.02)	0.04 (0.02)	0.01 (0.02)	0.04 (0.02)		0.02 (0.02)	0.02 (0.02)		-0.01 (0.02)	-0.01 (0.02)
Province: Eastern Cape	0.03 (0.04)		0.01 (0.01)	0.02 (0.02)	0.01 (0.01)		-0.04 (0.01)	-0.01 (0.02)	-0.02 (0.02)	-0.02 (0.01)	-0.03 (0.02)	
Province: Free State	0.06 (0.03)	0.04 (0.02)		0.02 (0.02)			-0.02 (0.01)		0.02 (0.01)	0.01 (0.01)	0.00 (0.02)	
Province: Gauteng	0.12 (0.06)	0.02 (0.02)	-0.01 (0.02)	-0.01 (0.02)		0.02 (0.02)	0.01 (0.02)	0.02 (0.02)			0.02 (0.02)	0.04 (0.02)
Province: KwaZulu-Natal	0.08 (0.05)	0.04 (0.02)		0.03 (0.02)						0.01 (0.02)		-0.02 (0.02)
Province: Limpopo	0.08 (0.04)	0.03 (0.02)	0.01 (0.02)	-0.02 (0.02)	-0.01 (0.02)					0.02 (0.02)	0.02 (0.02)	
Province: Mpumalanga	0.06 (0.04)	0.01 (0.02)	-0.02 (0.02)		-0.04 (0.02)				0.02 (0.02)			0.03 (0.02)
Province: North West	0.04 (0.04)	0.04 (0.02)		-0.04 (0.02)						-0.01 (0.02)	0.03 (0.02)	
Province: Western Cape	0.05 (0.04)		-0.02 (0.02)		0.02 (0.01)		-0.03 (0.02)	-0.02 (0.02)	-0.03 (0.02)	-0.05 (0.01)	-0.01 (0.02)	0.00 (0.02)
Challenge: DNK where	0.04 (0.02)	0.02 (0.02)		-0.01 (0.01)					-0.01 (0.01)		0.02 (0.01)	0.01 (0.02)
Challenge: Too expensive	0.05 (0.02)	0.00 (0.02)	0.02 (0.01)		0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.03 (0.02)				0.02 (0.02)
Challenge: Too time consuming	0.02 (0.02)				-0.02 (0.01)					0.01 (0.01)	-0.01 (0.02)	0.04 (0.02)
Challenge: Mistrust fact-checkers	0.02 (0.02)	-0.01 (0.02)	0.01 (0.01)	0.01 (0.02)	-0.02 (0.01)		-0.02 (0.01)		0.03 (0.01)	-0.01 (0.01)		0.00 (0.01)
Challenge: Fact-checks are irrelevant	0.04 (0.02)	-0.02 (0.02)	0.03 (0.01)					0.03 (0.02)		0.01 (0.01)	0.02 (0.02)	
Get news from: Other social media	0.01 (0.02)	0.02 (0.02)	-0.01 (0.02)	0.03 (0.02)			0.06 (0.01)	0.04 (0.02)				-0.03 (0.02)
Get news from: Family	0.02 (0.02)	0.07 (0.02)	-0.04 (0.02)	0.03 (0.02)	0.02 (0.02)			0.06 (0.02)	0.03 (0.01)			-0.04 (0.02)
Get news from: Radio/TV	0.05 (0.02)	0.03 (0.02)	0.04 (0.02)	0.03 (0.02)	0.06 (0.01)	0.02 (0.02)			0.02 (0.02)	-0.02 (0.01)	0.05 (0.02)	0.05 (0.02)
Get news from: WhatsApp	-0.02 (0.02)	0.01 (0.02)	-0.01 (0.02)	0.01 (0.02)		-0.04 (0.02)		-0.01 (0.02)		0.02 (0.02)	-0.04 (0.02)	0.02 (0.02)
WA news freq: Family	0.02 (0.02)	-0.03 (0.02)	0.00 (0.02)	-0.06 (0.02)	-0.03 (0.02)	-0.01 (0.02)				0.03 (0.02)		0.06 (0.02)
WA news freq: Large groups	0.06 (0.02)	0.02 (0.02)	0.03 (0.02)			-0.02 (0.02)	0.02 (0.02)		0.04 (0.02)	0.08 (0.02)		0.00 (0.02)
WA news freq: Orgs	0.04 (0.02)	-0.03 (0.02)	0.09 (0.02)		-0.01 (0.02)	-0.01 (0.02)	0.03 (0.02)	0.01 (0.02)	0.03 (0.02)	0.04 (0.02)		0.07 (0.02)
Behavior: Stayed home	0.02 (0.02)		0.01 (0.02)		-0.05 (0.02)			-0.04 (0.02)	0.02 (0.02)	-0.03 (0.01)	0.02 (0.02)	
Behavior: Visited indoors (reversed)	0.01 (0.02)	-0.01 (0.02)	0.03 (0.01)	0.01 (0.02)	0.03 (0.02)	0.03 (0.02)		-0.01 (0.02)			0.04 (0.02)	0.00 (0.01)
Behavior: Wore mask	-0.03 (0.02)	-0.04 (0.02)	0.02 (0.01)		-0.01 (0.01)		0.02 (0.01)	-0.03 (0.02)		-0.02 (0.01)	0.00 (0.02)	
How often listens to podcasts	0.18 (0.02)	0.01 (0.02)	0.03 (0.02)				0.02 (0.02)	0.00 (0.02)	0.06 (0.02)	0.05 (0.02)	0.01 (0.02)	0.05 (0.02)
Listens to WCW	0.09 (0.02)	0.02 (0.02)	-0.03 (0.02)		0.03 (0.02)			0.03 (0.02)	0.02 (0.01)	0.05 (0.02)	-0.05 (0.01)	0.05 (0.02)
Listens to Money Show	-0.01 (0.02)	-0.05 (0.02)	0.01 (0.01)	-0.03 (0.01)	-0.04 (0.02)	-0.03 (0.02)		0.06 (0.02)		0.02 (0.02)		0.02 (0.02)
Fake news leads to harmful behavior	0.05 (0.02)	0.01 (0.02)	0.03 (0.01)	-0.02 (0.02)				0.01 (0.02)		0.01 (0.01)		0.07 (0.02)
Trust on WA: Family members	0.00 (0.02)	-0.08 (0.02)	0.01 (0.02)	-0.07 (0.02)	-0.09 (0.02)	-0.03 (0.02)	0.08 (0.02)	0.01 (0.02)	0.00 (0.02)	0.08 (0.02)	-0.03 (0.02)	0.03 (0.02)
Trust: Info from other social media	-0.01 (0.02)		0.00 (0.02)	-0.05 (0.02)	-0.04 (0.02)	-0.01 (0.02)	0.15 (0.03)	0.04 (0.02)		0.03 (0.02)		0.03 (0.02)
Trust on WA: Orgs	0.03 (0.02)	0.05 (0.02)	0.05 (0.02)	0.03 (0.02)	0.02 (0.02)					-0.03 (0.02)	0.01 (0.02)	0.03 (0.02)
Trust: Info from radio/TV	0.02 (0.02)	0.04 (0.02)		0.04 (0.02)	0.05 (0.02)	0.02 (0.02)			0.02 (0.02)	0.02 (0.02)	0.03 (0.02)	0.08 (0.02)
Trust: Info from WhatsApp	0.01 (0.02)	-0.06 (0.02)	0.04 (0.02)		-0.04 (0.02)	-0.02 (0.02)				0.03 (0.02)	0.03 (0.02)	0.06 (0.02)
Trust on WA: Large groups	0.01 (0.02)	0.05 (0.02)	-0.05 (0.02)		0.03 (0.02)			0.02 (0.02)	0.02 (0.02)	0.01 (0.02)		-0.02 (0.02)
How often verify	0.02 (0.02)	0.03 (0.02)	0.06 (0.01)	0.03 (0.02)	0.02 (0.02)	0.05 (0.02)			0.17 (0.02)	-0.03 (0.01)	0.01 (0.02)	0.00 (0.02)
To verify: Ask in person (reversed)	0.04 (0.02)	0.02 (0.02)				0.04 (0.02)		-0.01 (0.02)	0.03 (0.02)	0.01 (0.01)	0.02 (0.02)	0.00 (0.02)
To verify: Ask family on WA (reversed)	0.00 (0.02)	0.00 (0.02)	0.03 (0.01)			0.03 (0.02)		-0.03 (0.02)		-0.01 (0.01)	-0.02 (0.02)	
To verify: Ask others on WA (reversed)	0.01 (0.02)	0.00 (0.02)	-0.01 (0.01)		0.05 (0.02)	0.02 (0.02)				-0.04 (0.01)	0.01 (0.02)	-0.05 (0.02)
To verify: Use fact-checker	0.01 (0.02)	0.07 (0.02)	-0.01 (0.01)	0.03 (0.02)	0.05 (0.01)	0.07 (0.02)	-0.05 (0.01)		0.04 (0.01)	-0.05 (0.01)	0.05 (0.01)	0.00 (0.02)
To verify: Post on social media (reversed)	-0.04 (0.02)	0.00 (0.02)	-0.03 (0.01)	0.01 (0.01)	0.01 (0.01)	0.05 (0.02)	-0.08 (0.01)	-0.03 (0.02)		-0.02 (0.01)	0.01 (0.02)	
To verify: Use internet	0.01 (0.02)	0.09 (0.02)	-0.02 (0.01)	0.07 (0.02)	0.07 (0.02)	0.08 (0.02)		0.02 (0.02)	0.04 (0.02)	-0.03 (0.01)	0.03 (0.02)	-0.03 (0.02)
Locality: Urban		-0.01 (0.04)	-0.02 (0.04)		0.01 (0.04)	0.01 (0.04)	0.06 (0.04)			-0.03 (0.04)	0.03 (0.04)	
Age: 35-44		-0.02 (0.03)	0.03 (0.01)	-0.02 (0.03)						-0.05 (0.02)	-0.03 (0.03)	
Trust: Info from family		0.02 (0.02)	0.00 (0.02)	0.04 (0.02)	0.02 (0.02)	0.02 (0.02)		-0.06 (0.02)			0.02 (0.02)	-0.02 (0.02)
ICW: Trust social media							0.07 (0.03)					
ICW: Consume social media								0.14 (0.02)				
ICW: Sharing										0.15 (0.02)		
ICW: COVID-19 beliefs and behavior											0.16 (0.02)	
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Control SD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
R^2	0.23	0.27	0.14	0.13	0.16	0.10	0.17	0.14	0.14	0.24	0.14	0.20
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: Regressions of ICW index outcomes used in main figures in text including all LASSO-selected controls. Column header provides Figure corresponding to relevant ICW outcome in the manuscript. All specifications are estimated using OLS, and adjust for randomization block fixed effects. Heteroskedasticity-robust standard errors in parentheses.

Table F15: Disaggregated estimation ICW outcomes (including LASSO-selected controls)

Figure	3a	3b	3c	4a	4b	5a	5b	6a	6b	6c	7a	7b
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Placebo incentives	0.32 (0.05)	0.13 (0.05)	0.06 (0.05)	0.06 (0.05)	0.00 (0.05)	0.05 (0.05)	-0.04 (0.05)	-0.02 (0.05)	-0.04 (0.05)	0.00 (0.04)	-0.03 (0.05)	0.09 (0.05)
Text information	-0.01 (0.06)	0.34 (0.06)	0.23 (0.06)	0.12 (0.06)	0.11 (0.06)	0.17 (0.06)	-0.14 (0.06)	-0.07 (0.06)	-0.13 (0.06)	-0.10 (0.05)	0.15 (0.06)	0.08 (0.06)
Short podcast	0.64 (0.05)	0.38 (0.05)	0.24 (0.04)	0.02 (0.05)	0.04 (0.04)	0.12 (0.05)	-0.02 (0.04)	0.02 (0.04)	-0.04 (0.04)	0.02 (0.04)	0.02 (0.04)	0.11 (0.04)
Long podcast	0.65 (0.05)	0.39 (0.05)	0.17 (0.04)	0.00 (0.05)	0.13 (0.04)	0.03 (0.05)	-0.07 (0.04)	0.01 (0.04)	0.02 (0.04)	-0.01 (0.04)	-0.02 (0.05)	0.03 (0.05)
Empathetic podcast	0.65 (0.05)	0.50 (0.05)	0.20 (0.04)	0.14 (0.05)	0.16 (0.04)	0.11 (0.05)	-0.14 (0.04)	-0.03 (0.04)	-0.05 (0.05)	-0.07 (0.04)	-0.06 (0.04)	0.02 (0.05)
Gender: Female	-0.06 (0.04)	-0.03 (0.04)	-0.01 (0.03)	-0.03 (0.04)		-0.01 (0.04)		0.01 (0.03)			0.11 (0.04)	-0.02 (0.04)
Locality: Peri-urban	0.01 (0.03)	0.11 (0.06)		0.13 (0.05)			-0.01 (0.02)	0.01 (0.03)				
Locality: Rural	0.06 (0.03)	0.08 (0.06)	0.00 (0.02)	0.08 (0.05)		0.00 (0.02)		-0.02 (0.03)		-0.02 (0.02)		0.06 (0.02)
Age: 18-24	0.10 (0.04)		-0.09 (0.02)	-0.03 (0.02)	-0.21 (0.02)	0.03 (0.02)		-0.05 (0.02)	-0.05 (0.02)	-0.02 (0.02)	-0.17 (0.04)	0.22 (0.03)
Age: 25-34	0.13 (0.04)	0.01 (0.02)		-0.04 (0.02)	-0.11 (0.02)	0.01 (0.02)					-0.10 (0.04)	0.17 (0.03)
Age: 35-44	0.03 (0.03)	-0.02 (0.02)	0.03 (0.01)							-0.05 (0.02)	-0.03 (0.03)	0.04 (0.02)
Education: Secondary	0.02 (0.02)	0.05 (0.01)	-0.01 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.02)	-0.02 (0.01)			-0.02 (0.01)		-0.02 (0.01)
Education: University	0.03 (0.02)	0.06 (0.02)	-0.06 (0.02)	0.04 (0.02)	0.01 (0.02)	0.04 (0.02)		0.02 (0.02)	0.02 (0.02)		-0.01 (0.02)	
Province: Eastern Cape	-0.03 (0.02)		0.01 (0.01)	0.02 (0.02)	0.01 (0.01)	-0.01 (0.02)	-0.04 (0.01)	-0.01 (0.02)	-0.02 (0.02)	-0.02 (0.01)	-0.03 (0.02)	
Province: KwaZulu-Natal	0.00 (0.02)	0.04 (0.02)		0.03 (0.02)						0.01 (0.02)		-0.02 (0.02)
Province: Limpopo	0.01 (0.02)	0.03 (0.02)	0.01 (0.02)	-0.02 (0.02)	-0.01 (0.02)					0.02 (0.02)	0.03 (0.02)	
Province: Mpumalanga	-0.01 (0.02)	0.01 (0.02)	-0.02 (0.02)	0.00 (0.02)	-0.03 (0.02)				0.02 (0.02)			0.03 (0.02)
Province: North West	-0.03 (0.02)	0.04 (0.02)		-0.03 (0.02)						-0.01 (0.02)	0.03 (0.02)	
Province: Western Cape	-0.03 (0.02)		-0.02 (0.02)		0.02 (0.01)		-0.03 (0.02)	-0.03 (0.02)	-0.03 (0.02)	-0.05 (0.01)	-0.01 (0.02)	0.00 (0.02)
Challenge: DNK where	0.04 (0.02)	0.02 (0.02)	0.01 (0.01)	-0.01 (0.01)				-0.01 (0.02)	-0.01 (0.01)		0.02 (0.01)	0.02 (0.02)
Challenge: Too expensive	0.05 (0.02)	0.01 (0.02)	0.02 (0.01)		0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.03 (0.02)				0.02 (0.02)
Challenge: Too time consuming	0.02 (0.02)				-0.02 (0.01)					0.01 (0.01)	-0.01 (0.02)	0.04 (0.02)
Challenge: Mistrust fact-checkers	0.03 (0.02)	-0.01 (0.02)	0.01 (0.01)	0.01 (0.02)	-0.02 (0.01)		-0.02 (0.01)		0.03 (0.01)	-0.01 (0.01)		0.00 (0.01)
Challenge: Fact-checks are irrelevant	0.04 (0.02)		0.03 (0.01)			-0.01 (0.02)		0.03 (0.02)		0.02 (0.01)	0.02 (0.02)	
Get news from: Family	0.02 (0.02)	0.07 (0.02)	-0.04 (0.02)	0.03 (0.02)	0.02 (0.02)	0.02 (0.02)		0.07 (0.02)	0.03 (0.01)			-0.04 (0.02)
Get news from: Radio/TV	0.05 (0.02)	0.03 (0.02)	0.04 (0.02)	0.03 (0.02)	0.06 (0.01)	0.02 (0.02)		-0.02 (0.02)	0.02 (0.02)	-0.01 (0.01)	0.04 (0.02)	0.05 (0.02)
WA news freq: Family	0.02 (0.02)	-0.04 (0.02)	0.00 (0.02)	-0.06 (0.02)	-0.03 (0.02)	-0.01 (0.02)				0.03 (0.02)	0.01 (0.02)	0.07 (0.02)
WA news freq: Large groups	0.05 (0.02)	0.02 (0.02)	0.03 (0.02)			-0.03 (0.02)	0.02 (0.02)		0.04 (0.02)	0.08 (0.02)		0.00 (0.02)
WA news freq: Orgs	0.03 (0.02)	-0.03 (0.02)	0.08 (0.02)	0.00 (0.02)	0.00 (0.02)	-0.01 (0.02)	0.03 (0.02)	0.01 (0.02)	0.03 (0.02)	0.04 (0.02)		0.07 (0.02)
Behavior: Stayed home	0.02 (0.02)		0.01 (0.02)		-0.05 (0.02)			-0.04 (0.02)	0.02 (0.02)	-0.03 (0.01)	0.02 (0.02)	
Behavior: Wore mask	-0.03 (0.02)	-0.04 (0.02)	0.02 (0.01)	0.00 (0.02)	0.00 (0.01)		0.02 (0.01)	-0.03 (0.02)		-0.02 (0.01)	0.00 (0.02)	
How often listens to podcasts	0.18 (0.02)	0.01 (0.02)	0.03 (0.02)				0.02 (0.02)	0.00 (0.02)	0.06 (0.02)	0.05 (0.02)	0.01 (0.02)	0.05 (0.02)
Listens to WCW	0.08 (0.02)	0.02 (0.02)	-0.03 (0.02)		0.03 (0.02)			0.03 (0.02)	0.02 (0.01)	0.05 (0.02)	-0.05 (0.01)	0.05 (0.02)
Fake news leads to harmful behavior	0.04 (0.02)	0.01 (0.02)	0.03 (0.01)	-0.02 (0.02)				0.01 (0.02)		0.01 (0.01)		0.07 (0.02)
Trust: Info from other social media	-0.01 (0.02)		0.00 (0.02)	-0.05 (0.02)	-0.04 (0.02)	-0.01 (0.02)	0.15 (0.03)	0.04 (0.02)		0.04 (0.02)		0.03 (0.02)
Trust on WA: Orgs	0.03 (0.02)	0.05 (0.02)	0.05 (0.02)	0.03 (0.02)	0.02 (0.02)					-0.03 (0.02)	0.01 (0.02)	0.03 (0.02)
Trust: Info from radio/TV	0.02 (0.02)	0.04 (0.02)		0.04 (0.02)	0.05 (0.02)	0.02 (0.02)			0.02 (0.02)	0.02 (0.02)	0.03 (0.02)	0.08 (0.02)
Trust: Info from WhatsApp	0.01 (0.02)	-0.06 (0.02)	0.04 (0.02)	-0.02 (0.02)	-0.04 (0.02)	-0.02 (0.02)				0.03 (0.02)	0.03 (0.02)	0.06 (0.02)
Trust on WA: Large groups	0.01 (0.02)	0.05 (0.02)	-0.05 (0.02)		0.03 (0.02)		0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.01 (0.02)	-0.02 (0.02)	-0.02 (0.02)
How often verify	0.01 (0.02)	0.03 (0.02)	0.06 (0.01)	0.03 (0.02)	0.02 (0.02)	0.05 (0.02)			0.17 (0.02)	-0.04 (0.01)	0.01 (0.02)	
To verify: Ask in person (reversed)	0.04 (0.02)	0.02 (0.02)				0.04 (0.02)		-0.02 (0.02)	0.03 (0.02)	0.01 (0.01)	0.02 (0.02)	0.00 (0.02)
To verify: Ask others on WA (reversed)	0.01 (0.02)		-0.01 (0.01)		0.04 (0.02)	0.02 (0.02)				-0.04 (0.01)	0.01 (0.02)	-0.05 (0.02)
To verify: Use fact-checker	0.01 (0.02)	0.06 (0.02)	-0.01 (0.01)	0.03 (0.02)	0.05 (0.01)	0.07 (0.02)	-0.05 (0.01)		0.04 (0.01)	-0.05 (0.01)	0.05 (0.01)	0.00 (0.02)
To verify: Post on social media (reversed)	-0.04 (0.02)		-0.03 (0.01)	0.01 (0.01)	0.01 (0.01)	0.05 (0.02)	-0.08 (0.01)	-0.03 (0.02)		-0.02 (0.01)	0.01 (0.02)	
Locality: Urban		0.08 (0.08)	-0.02 (0.04)	0.09 (0.07)	0.00 (0.04)	0.01 (0.04)	0.06 (0.04)			-0.03 (0.04)	0.03 (0.04)	
Province: Free State		0.04 (0.02)		0.02 (0.02)		0.02 (0.02)	-0.02 (0.01)		0.02 (0.01)	0.01 (0.01)	0.00 (0.02)	
Province: Gauteng		0.02 (0.02)	-0.01 (0.02)	-0.01 (0.02)		0.02 (0.02)	0.01 (0.02)	0.02 (0.02)			0.02 (0.02)	0.04 (0.02)
Get news from: Other social media		0.02 (0.02)	-0.01 (0.02)	0.03 (0.02)			0.06 (0.01)	0.04 (0.02)				-0.03 (0.02)
Get news from: WhatsApp		0.01 (0.02)	-0.01 (0.02)	0.01 (0.02)		-0.04 (0.02)		-0.01 (0.02)		0.02 (0.02)	-0.05 (0.02)	0.02 (0.02)
Behavior: Visited indoors (reversed)		-0.01 (0.02)	0.03 (0.01)	0.01 (0.02)	0.03 (0.02)	0.03 (0.02)		-0.01 (0.02)			0.04 (0.02)	0.00 (0.01)
Listens to Money Show		-0.05 (0.02)	0.01 (0.01)	-0.03 (0.01)	-0.04 (0.02)	-0.03 (0.02)		0.06 (0.02)		0.02 (0.02)	0.01 (0.02)	0.02 (0.02)
Trust: Info from family		0.02 (0.02)	0.00 (0.02)	0.04 (0.02)	0.02 (0.02)	0.02 (0.02)		-0.06 (0.02)			0.03 (0.02)	-0.02 (0.02)
Trust on WA: Family members		-0.08 (0.02)	0.01 (0.02)	-0.06 (0.02)	-0.09 (0.02)	-0.03 (0.02)	0.07 (0.02)	0.01 (0.02)	0.00 (0.02)	0.08 (0.02)	-0.03 (0.02)	0.03 (0.02)
To verify: Ask family on WA (reversed)		0.00 (0.02)	0.03 (0.01)			0.03 (0.02)		-0.03 (0.02)		-0.01 (0.01)	-0.02 (0.02)	
To verify: Use internet		0.09 (0.02)	-0.02 (0.01)	0.08 (0.02)	0.07 (0.02)	0.08 (0.02)		0.02 (0.02)	0.04 (0.02)	-0.03 (0.01)	0.03 (0.02)	-0.04 (0.02)
ICW: Trust social media							0.06 (0.03)					
ICW: Consume social media								0.14 (0.02)				
ICW: Sharing										0.15 (0.02)		
ICW: COVID-19 beliefs and behavior											0.16 (0.02)	
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Control SD	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
R^2	0.25	0.27	0.14	0.13	0.16	0.11	0.18	0.14	0.14	0.24	0.14	0.20
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: Regressions of ICW index outcomes used in main figures in text including all LASSO-selected controls. Column header provides Figure corresponding to relevant ICW outcome in the manuscript. All specifications are estimated using OLS, and adjust for randomization block fixed effects. Heteroskedasticity-robust standard errors in parentheses.

G Pre-analysis Plan

Can fact-checking podcasts combat misinformation in South Africa?

Potentially harmful misinformation runs rampant on social media across a wide set of countries. We explore how fact-checking pod- casts can be used to inhibit citizens' exposure to misinformation, increase their knowledge about COVID-19, and ultimately increase their compliance with public health policies. The intervention we study uses WhatsApp-delivered podcasts as an attention-catching method of delivering verified information to individuals who may otherwise have limited access to credible online sources. We partner with the first and largest fact-checking organization in sub-Saharan Africa, Africa Check, and randomize the delivery of variants of their programming to a recruited sample of participants in a panel survey in South Africa. The study has implications both for understanding how citizens' exposure to misinformation can be reduced with low- cost interventions and how the correction of false information can increase citizens' trust in public policies.

1 Introduction

Misinformation about social, political, and public health issues is a growing problem in many sub-Saharan African countries, where the rapid spread of social media technologies has led to the increasingly viral spread of misinformation (Zarocostas 2020). The COVID-19 crisis, for example, has highlighted the need to identify ways to counter social media posts spreading fake cures, false information about vaccines, and other misinformation (Van Bavel et al. 2020). In particular, the spread of misinformation through WhatsApp has become a major challenge, since high data costs for Internet access mean that discounted WhatsApp data bundles are the only affordable source of online information for many people in southern Africa (The Economist 2019). Moreover, since WhatsApp, unlike other social networks like Facebook or Twitter, is protected by end-to-end encryption, misinformation can spread while remaining especially difficult to monitor and regulate. The rise of misinformation is concerning because it may cause individuals to make harmful choices, whether by inducing false beliefs, priming particular modes of thinking, or by crowding out more credible information.

As social media is cost efficient for citizens in developing country settings, our project seeks to counter misinformation through these same popular low-cost channels. We propose to test the effectiveness of a WhatsApp-delivered fact-checking biweekly podcast on knowledge, attitudes, and behavior related to controversial topics which have been the subject of viral misinformation. We are interested in studying the longer-term effects of exposure to misinformation-targeting interventions, with a view toward understanding how to inoculate news consumers from believing potentially harmful, unverified information. To the extent that citizens seek to form accurate beliefs, rather than engage in motivated reasoning or adopt views of which they doubt the credibility, our intervention is expected to alter how citizens process information, what they believe, and potentially how they behave.

1.1 Literature

There is a growing literature on the efficacy of policies that combat fake news and viral misinformation, including (but not limited to) fact-checking interventions (Nyhan 2020; Pennycook et al. 2021). Most commonly, researchers provide corrective information to sample surveys and measure whether such researcher-provider information can shift knowledge and opinions about related topics in surveys. On average, studies in this literature demonstrate that it is possible to increase the accuracy of participants' beliefs through fact-checks, although effects vary depending on participants' prior beliefs and knowledge (Walter et al. 2020).

However, most fact-checking studies to date have important limitations. One challenge is that many survey-based fact-check experiments control the respondent's information environment for a short study period, raising the salience of researcher-provided fact-checks (Brashier et al. 2021; Guess et al. 2020). However in real life individuals can choose from multiple competing sources of information to consume or ignore. These experiments are also limited by the short time between provision of corrective information and survey implementation. By contrast, this study will use a field experiment in which information is provided naturalistically to respondents over a 6 month period; they are modestly incentivized to consume this information but can also choose to ignore it if they prefer.

In addition, the experimental design aims to test several mechanisms, suggested by both theory and the existing literature, which are hypothesized to strengthen the value of the fact-checking treatments. First, we focus on the role of emotion. A large literature demonstrates that belief in fake news (Martel, Pennycook and Rand 2020), as well as updating beliefs based on fact-checks (Gaines et al. 2007), is not a purely rational cognitive process—rather, it is deeply shaped by the emotional and identity commitments of individuals (Jerit and Zhao 2020). To date, the literature on emotions and fact-checking has largely focused on how negative or partisan emotions, either inadvertently or purposefully elicited by fact-checking treatments, reduce the ability of individuals to update and learn (Van Bavel and Pereira 2018). We add to these studies by examining another form of emotion—specifically, an appeal to the broader social good—as a way to elicit greater levels of updating. Another area of uncertainty in the literature relates to the length and complexity of fact-check messages. While meta-analysis of fact-check length on outcomes suggests no impact (Walter et al. 2020), we are not aware of evidence on the length of audio content (such as podcasts) or contrasts of text-based to audio-based interventions.

1.2 Intervention

The intervention we study consists of a set of informational treatments administered through WhatsApp. For each of these, we collaborate with the South Africa-based civil society organization Africa Check—the first and largest fact-checking organization in sub-Saharan Africa. As part of Africa Check's programming, the organization partnered with Volume, an independent South African podcasting firm, to launch "What's Crap on WhatsApp?" (WCW), a short biweekly podcast which debunks locally-relevant misinformation. Episodes generally last 6-8 minutes and cover three specific stories which have circulated on social media in South Africa in the preceding few weeks, with items occasionally suggested by podcast subscribers.

The podcast is disseminated to subscribers directly through WhatsApp, and consumes relatively little data to download. Relative to other misinformation-targeting interventions, the podcast has two potential advantages. First, it is a professionally-produced product, and are therefore likely to be more accessible, entertaining, and engaging than more anodyne modes of information delivery. Second, due to its mode of delivery through WhatsApp, it potentially allows listeners to quickly share content with their contacts, offering a chance for corrective information to spread relatively quickly within users' social networks. Our study experimentally tests the impact of the podcast intervention. Further, as detailed below, we produce three variants of each podcast episode—the normal version that Africa Check already disseminates to its subscribers, a version that seeks to empathize with participants that might have been fooled by the misinformation that the podcast shows to be fake, and a shortened version—and its accompanying messaging in order to understand which aspects of the intervention drive its potential effects. We further compare the podcasts with a simpler text-based intervention that only conveys the results of fact-checks via the basic WhatsApp message received by all participants that also receive the podcast.

Online recruitment for the study commenced in October 2020 and continues at the time of writing. This pre-analysis plan was submitted after the earliest batch of participants took the endline survey (n=126) but prior to any endline data analysis.

2 Research design

This section provides an overview of our study sample recruitment, treatment variants and randomization, data collection, and estimation strategy.

2.1 Sample recruitment and baseline survey

Individuals are eligible for study participation if they are currently living in South Africa, have a South African phone number, are at least 18 years of age, and are WhatsApp users. We recruit our study sample using a set of Facebook ads (see Appendix A for a sample ad). In an effort to ensure reasonably broad geographical coverage, we stratify these ads at the province-gender-age level, generating a total of 36 different ads.[1] The ads invite participation in a research study relating to social media in South Africa for which participants will be provided airtime.

Upon clicking an ad, potential participants are first redirected to a Qualtrics landing page where they read the informed consent information and agree to participate. If the participant agrees to proceed, they are then asked to send a WhatsApp message to the phone number associated with our interactive project WhatsApp chatbot. The chatbot repeats the informed consent process and further determines eligibility based on demographic information that the participant provides at the start of the baseline survey.[2]

Conditional on eligibility, the chatbot then immediately administers the baseline survey instrument. The baseline survey records (1) initial attitudes on different sources of information, both off- and online; (2) attitudes and behaviors regarding misinformation and fact-checking; (3) baseline knowledge about current affairs and COVID-19; (4) podcast listening habits; (5) behaviors relating to social distancing measures that were undertaken at the start of the pandemic in South Africa. As part of the baseline survey, participants are required to send a WhatsApp message to a phone number run by Africa Check and add that number to their phone contacts,[3] which we validate. They are informed that, subsequent to the baseline survey, Africa Check might send them information. Participants are incentivized with R30 (approximately $2) in mobile airtime credit for

[1] Specifically, ads are targeted according to (1) province of the user, of which there are 9 total (2) whether the user is male or female (3) whether they are between 18-29 or 30-49 years old. Our pilot testing suggested that attracting over-50s to participate in the study was extremely expensive.

[2] Potential participants found to be ineligible have their phone numbers banned by the chatbot to avoid falsified eligibility information. See Appendix B for an example of the chatbot interface.

[3] This is required for Africa Check to be able to send them their podcast through a WhatsApp list.

completing the baseline survey and for successfully messaging Africa Check's WhatsApp account.

2.2 Random assignment and experimental treatments

Due to the rolling nature of study recruitment (detailed below), we block randomize batches of participants into treatment conditions once every two weeks. We block on a set of variables including demographic characteristics, social media usage, attitudes towards different media sources, and knowledge regarding pieces of misinformation.[4]

We adopt a "nested" blocking strategy, whereby we construct two levels of concentric randomization blocks. At the lower level, a block contains 19 respondents. To account for the possibility of attrition reducing within-block variation in treatment assignment, we also aggregate these blocks into higher-level blocks containing a greater number of participants—specifically, the larger blocks combine two smaller blocks to contain 38 individuals. As a result, with a choice of blocks defined at different levels of granularity, for estimation purposes we will be able to choose the level which minimizes within-block participant characteristic variation subject to sufficient levels of endline survey completion across the different treatment conditions within a block.

Table 1: Treatment Assignment

		Text only		Short podcast (3-5 min)		Long podcast (6-8 min)		Emotional podcast (6-8 min)	
	Control	F	S	F	S	F	S	F	S
Podcast incentives	0.00	0.04	0.04	0.09	0.09	0.09	0.09	0.09	0.09
Placebo incentives	0.24	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02

Table presents the sample sizes of the planned design. 'F': factual message; 'S': social message. All podcast treatments also include the text message via WhatsApp.

Study participants are randomly assigned to either *control* or one treatment group. The treatments are distinguished along three dimensions: (1) mode of information delivery; (2) messaging encouraging information consumption; (3) whether participants are incentivized to take up the treatment. Table 1 summarizes the research design with the approximate share of participants assigned to each cell. In total, we are targeting a baseline survey sample of around 5,500 participants, with the expectation of approximately 2,000 completing the full six month study. In Appendix C and D, we provide a sample script of the different messaging and quizzes, respectively.

[4]We use the `R` package `blocktools` to assign blocks, batch by batch, based on a greedy algorithm using Mahalanobis distance.

2.2.1 Mode of information delivery

First, we vary how the information contained in the podcasts is delivered to participants. We administer four treatment variants: (1) a text-only treatment, (2) a short podcast, (3) a longer podcast, and (4) a longer podcast which includes emotional appeals. Each variant contains the same core information regarding the truthfulness of (often viral) news fact-checked by Africa Check; the variation comes from the mode of information delivery.

The text-only treatment contains a true, false, or misleading tag for three pieces of news that Africa Check has identified for the specific week. This information is summarized simply in a single sentence. Each such WhatsApp message also includes a link to a longer article on Africa Check's website for each item. The items that WCW covers are generally sourced from social media, are mostly shown to be false, and frequently cover issues relating to public health, government, and immigration.

The text-only fact-checking content is contrasted with three more engaging, but also more time-consuming, forms of information dissemination via a podcast. Each form of the podcast is sent as part of a WhatsApp message that also contains the text-only information; the podcast thus come in addition to the text-only treatment. The short podcast is a 3-5 minute conversation between the man and woman serving as co-hosts, explaining, discussing, and evaluating the truth of the same three pieces of viral news. The short conversation of each viral news items culminates in concluding whether it is true, false, or misleading, and how Africa Check came to that conclusion. The longer podcast is a 6-8 minute conversation between the co-hosts. In the longer podcast, the co-hosts go into greater depth about the sources that they consulted and the conclusions they are able to draw. In the emotional variant of the longer podcast, which also lasts for around 6-8 minutes, the hosts specifically acknowledge in an empathetic manner the underlying reasons—such as economic insecurity or distrust in the state—which might lead people to be susceptible to a particular piece of misinformation. The rationale is that by acknowledging the emotions behind misinformation, this variant of the treatment may increase engagement with the podcast and information, especially among those fooled by the misinformation who may be more likely to engage in motivated reasoning. It may also increase the salience of fake news and fact-based decision making among listeners. However, since the emotional component is only added to one of the three fact checks in each episode, this treatment is relatively subtle.

2.2.2 Messaging encouraging information engagement

Along the second dimension, we vary the type of messaging used to induce participants to consume their informational treatment. Specifically, we vary whether participants receive a 'factual' WhatsApp message or a 'social' WhatsApp message. Under the 'factual' message condition, participants are sent a message which announces the availability of the podcast variant (or just contains the text variant summarizing the fact checks). Under the 'social' message variant, participants are sent the same message but containing an appeal which highlights the potential harms of misinformation—whether to participants' friends and family or society more broadly—and in some cases further emphasizes potential reputational benefits of being informed within a social network.

2.2.3 Incentivized treatment uptake

To maximize treatment uptake and continued engagement with the project (across mode of delivery, as well as in general), we further administer incentivized monthly quizzes that encourage participants to pay attention to the information provided. However, since the quizzes cover information from the treatment deliveries, incentivized quizzes can only be delivered to participants in treatment groups and not participants in the control group. Yet, not providing the control group with quizzes may introduce differential attrition. We therefore provide all participants with incentivized quizzes, but all control participants and a portion of treated participants are randomly assigned to receive "placebo" quizzes, which contain questions about pop culture or sports topics which are not covered in the treatment messages or podcasts. We specifically avoid political and current affairs topics for the placebo quizzes to minimize potential overlap with the content of the podcasts. We assign some treated participants to receive the placebo quizzes in order to test whether incentives are required for individuals to engage with the treatments.

Each quiz is six questions long and takes roughly two minutes to complete. If the participant answers less than four questions correctly, they receive R10; if they answer four or more questions correctly, they are rewarded with an additional R10 for a total of R20. These incentives are delivered in the form of mobile airtime credits. All participants are informed of which types of quiz questions they will receive at the outset of the study and their assignment is constant across quizzes.

2.3 Treatment delivery and data collection

Treatment delivery and data collection are all conducted through WhatsApp.

2.3.1 Treatment delivery

Once participants subscribe to the Africa Check WhatsApp account during the baseline survey, Africa Check assigns participants to a specific WhatsApp broadcast list associated with their treatment condition (or to no broadcast list for control). Then, Africa Check delivers the corresponding treatment combination to participants through messaging every two weeks.

2.3.2 Data collection

We collect survey data through the WhatsApp chatbot provider Landbot. Data is collected through the baseline survey, monthly quizzes, a midline survey administered three months into the study for a given batch, and finally an endline survey administered six months into the study for a given batch. Participants are enrolled on a rolling basis and are grouped into two-week "batches" to correspond with their biweekly treatment delivery from Africa Check. A sample of the study timeline is reproduced in Appendix E for each batch of participants. Quizzes contain material relevant to the two prior treatment deliveries.[5]

2.4 Estimation

To estimate the effect of treatment assignments on engagement with the fact-checking content and subsequent beliefs and behaviors, we use the midline and endline surveys (as well as the quiz answers) to compare treated individuals across different treatments conditions and with the

[5]For example, a podcast-incentivized quiz will ask participants quiz questions about content sent to participants in the preceding month; while a placebo-incentivized quiz will ask about pop culture events that occurred in the preceding month.

control condition. We start by describing the most general form of regression specification before then detailing how we will collapse treatment conditions to increase statistical power.

We estimate average treatment effects using the following OLS regression:

$$Y_{ib} = \alpha_b + \beta Y_{ib}^{pre} + \gamma \mathbf{X}_{ib}^{pre} + \boldsymbol{\tau} \mathbf{T}_{ib} + \varepsilon_{ib}, \tag{1}$$

where Y_{ib} is an outcome for respondent i from block b in a given survey wave, $\mathbf{T}_{ib}$ is the vector of individual treatment assignments, α_b are randomization block fixed effects,[6] Y_{ib}^{pre} is the baseline analog of the outcome (where feasible) and $\mathbf{X}_{ib}^{pre}$ is a vector of additional baseline covariates selected via LASSO.[7] The vector $\boldsymbol{\tau}$ captures the effect of each treatment condition; the effect of different treatment conditions can be identified by comparing elements within this vector. Robust (HC2) standard errors will be used throughout, except where survey waves are pooled (to examine quiz scores across treatment conditions and for questions repeated in midline and endline) when standard errors will be clustered at the individual level. We can further estimate heterogeneous and conditional treatment effects by pooling across relevant treatments and interacting $\mathbf{T}_{ib}$ in equation (1) with relevant predetermined covariates.

Although we can analyze each treatment condition separately, the study was designed with the intention of pooling across similar treatment conditions to increase statistical power. To examine how access to the fact-checking content by text-only messages and/or podcasts affect outcomes, we will pool across treatment conditions in the following ways:

1. Emotional podcast vs. long podcast vs. short podcast vs. text only vs. control: pool conditions across quiz incentives *and* across 'factual' and 'social' WhatsApp message types.

2. Long podcast vs. short podcast vs. text only vs. control: pool conditions across quiz incentives *and* across 'factual' and 'social' WhatsApp message types *and* across long and emotional podcasts.

3. Any podcast vs. text only vs. control: pool conditions across quiz incentives *and* across

[6] In practice we intend to report both of the potential blocking levels in our analyses.

[7] As potential covariates, we will consider all standardized baseline covariates and their interaction with $\mathbf{T}_{ib}$. For each outcome variable, we will use cross-validated LASSO to select the conditioning variables for inclusion in Equation (1). When examining heterogeneous effects, we will hold fixed the set of conditioning variables between estimating the ATE and the CATE.

'factual' and 'social' WhatsApp message types *and* across longer, shorter, and emotional podcasts.

4. Any fact-checking treatment vs. control: pool conditions across quiz incentives *and* across 'factual' and 'social' WhatsApp message types *and* across text only messages and all podcast types.

5. Differential effects of fact-checking treatments by encouragement message: pool conditions across quiz incentives.

6. Differential effects of fact-checking treatments by incentive: pool conditions across 'factual' and 'social' WhatsApp message types.

The first four of these comparisons constitute the analyses of principal interest. The fifth and sixth are important in conjunction with the engagement results (discussed next) for understanding whether any differences between treatment conditions reflect a greater probability of exposure to treatment across treatment conditions and/or differences in the content itself. For each type of analysis, we will report results that both include these observations in the control group and drop these observations from the analysis in the event that placebo incentives do not affect text only messages or podcast engagement.

To examine the effects of encouragement messages on engagement with the fact-checking content (which we measure in various ways described below), we will pool across treatment conditions in the following ways (excluding control group respondents that did not receive any content to engage with):

1. Factual vs. social encouragement messages crossed with podcast vs. placebo incentives, by fact-checking information type: no pooling.

2. Factual vs. social encouragement messages, by fact-checking information type: pool conditions across quiz incentives.

3. Factual vs. social encouragement messages, by any podcast vs. text only : pool conditions across quiz incentives *and* across all longer, shorter, and emotional podcast conditions.

4. Podcast vs. placebo incentives, by fact-checking information type: pool conditions across 'factual' and 'social' WhatsApp message types.

5. Podcast vs. placebo incentives, by any podcast vs. text only: pool conditions across 'factual' and 'social' WhatsApp message types *and* across all longer, shorter, and emotional podcast conditions.

2.4.1 Missing data

We expect to encounter two forms of missing data: attrition from surveys; and "don't know" responses to particular questions. To assess the extent to which differences in attrition across treatment conditions may introduce biases, we will: (i) use the equation specified above to examine the extent to which attrition varies across treatment groups; and (ii) compare balance tests of predetermined (baseline) covariates at the point of assignment (before attrition can occur) with balance tests among the non-attrited sample in the midline and endline surveys. In the event that we encounter severe attrition, we will seek to condition the sample on predetermined covariates for which there is limited imbalance and conduct analysis using Lee bounds. With regard to "don't know" responses to specific questions in a survey, such responses will be coded as "negatives"—that is to say, not doing the thing noted in the question (e.g. when asked about listening to podcasts "don't know" would be coded as "never", while for the importance of an issue "don't know" would be coded as "not at all important"); where "don't know" relates to a Likert scale, don't know will be coded as the median/neutral option (e.g. as "neither agree not disagree").

2.4.2 Low-quality responses

Low quality respondents are removed during the recruitment process using three attention-checking questions that randomly appear throughout the baseline survey. These attention-checking questions are designed such that they are easy to respond if respondents read the question (e.g. "What year is it?"). Respondents who do not pass these these questions are deemed ineligible to proceed with the study and are not included in the randomization process. Their phone numbers are also prevented from restarting the baseline survey.

Though we are able to ascertain a baseline level of response quality across all participants in the study using the aforementioned method, we further restrict the sample to conduct robustness checks in two ways. First, our own pilots of the baseline survey suggest that the entire survey cannot be plausibly comprehended and completed in less than 6 minutes. Therefore, as a conservative estimate, we conduct robustness checks using only the subsample of participants who took more than 8 minutes to complete either the baseline survey or endline surveys. Second, we obtain pre-treatment demographic data on the participant's province and level of education at baseline and midline. While it is possible that the participant may have moved during the study or may have attained additional education, such instances are likely to be rare. For a second set of robustness checks for data quality, we therefore restrict the sample only to individuals whose responses to these two questions match across baseline and midline.

2.4.3 Statistical inference

For hypotheses where we prespecify an expected direction, e.g. a positive effect of treatment on a given outcome, we will use one-sided t tests to evaluate the hypothesis. In the event that the coefficient has the opposite sign, we will use two-sided t tests to evaluate whether the null hypothesis can be rejected. Where no direction for a hypothesis is specified, we will instead conduct two-sided t tests.

3 Hypotheses

We next pre-specify our primary hypotheses by outcome family. For each family of outcomes, we also compute inverse covariance weighted (ICW) indices that are standardized relative to the control group.

The hypotheses below refer to the text only message and podcasts collectively as the treatment. However, across all hypotheses, we expect the effects of fact-checked information to be particularly concentrated among participants assigned to: (1) podcasts rather than text messages; (2) emotional podcasts rather than similarly-long non-emotional podcasts; (3) podcast-incentives rather than those assigned to placebo-incentives; (4) social messages rather than factual messages. For each of these predicted differences in effect magnitude, we conduct one-sided tests. We do not anticipate a particular direction for (5) longer podcasts rather than short podcasts, for which we conduct two-sided tests.

3.1 Exposure to intervention ("first stage")

We first expect that participants assigned to the treatment conditions should exhibit greater knowledge and awareness of the information they have received through the duration of the study at endline:

H1 : Access to fact-checking content increases exposure to, and knowledge about, information covered by the treatment deliveries.

We measure these effects using responses to questions about (1) participants' self-reported listening to podcasts, specifically WCW; (2) participants' correct answers to quizzes embedded in the midline and endline cover factual information from the two prior treatment deliveries; (3) the frequency with which participants report being alerted that particular pieces of information on social media are fake; (4) participants' knowledge about sources which can be used to verify information; (5) participants' knowledge about specific fact-checkers. In addition, we will combine core outcomes (1)-(3) using an ICW index; variables (4) and (5) will be analyzed separately because they are less direct measures of engagement. We can also compare the monthly podcast quiz scores between treatment conditions, but cannot draw comparisons with the group (or other treated groups) that only received the placebo quizzes.

3.2 Perceptions of misinformation and trust in information sources

We hypothesize that participants assigned to treatment should then become more aware of the extent of misinformation. In the context of our study, Africa Check debunks misleading or fake information that are shared on various social media websites through various friend and family networks. We therefore expect that:

H2 : Access to fact-checking content increases participants' perceptions of the extent of misinformation circulated through social media platforms.

We measure participants' perceptions of the extent of misinformation using: (1) participants' beliefs about how much information on platforms like WhatsApp, Facebook, and Twitter is false; and (2) how much information from WhatsApp groups (either consisting of close friends/family or large WhatsApp groups) is false. We will combine these two measures using an ICW index.

In addition to perceptions of the extent of misinformation, we also hypothesize that the treatment will induce a more general decrease in trust in information from the same set of sources:

H3 : Access to fact-checking content reduces participants' trust in information received on social media platforms.

We measure participants' trust in the information they receive from the same set of sources as H2, which we will similarly combine using an ICW index. We expect weaker treatment effects, if any, on beliefs about misinformation (and trust) relating to traditional media sources, such as radio, TV, and newspapers, which are generally more likely to verify the information they cover and are less frequently the targets of fact-checks on WCW.

3.3 Consumption and sharing behavior

We expect that the treatment, by shifting participants' beliefs about the credibility of different information sources, will change participants' behavior regarding consuming and sharing information:

H4 : Access to fact-checking content reduces participants' consumption, and sharing, of information from social media platforms.

H5 : Access to fact-checking content increases participants' attention to the veracity of information they encounter on social media platforms.

Specifically, for H4, we expect that treated participants will (1) consume less information from social media platforms (such as WhatsApp, Facebook, and Twitter) overall, and (2) more specifically from sources on WhatsApp aside from organizations to which they have subscribed. Additionally, due to their increased knowledge of the extent of misinformation, we expect that (3) treated participants in general should share and forward information on social media platforms less frequently. We will again combine these measures using an ICW index. We assess H5 based on responses to a set of questions about how much attention participants pay to the truthfulness of information they are sent on social media platforms.

3.4 Behavior around misinformation

A primary set of outcomes relates to participants' changes in behavior when presented with potential misinformation. We hypothesize that treatment will have the following effects on participants' behavior:

H6 : Access to fact-checking content changes participants' capacity to identify, and express skepticism on the basis of, characteristics of misinformation.

H7 : Access to fact-checking content changes participants' behavior in checking the veracity of information they encounter through social media platforms.

For H6, we primarily measure participants' beliefs about the characteristics of misinformation using a conjoint experiment embedded in the endline survey instrument. Across a set of four questions which hold fixed the truthfulness of a given claim (some of which are true and others are false), we vary whether participants are (1) provided a credible source for the claim; (2) told that the claim has been independently validated; (3) told that the piece of information was from a viral Facebook post; and (4) told that the claim came from a source that is likely to be subject to sensationalized fabrication. The potential importance of each characteristics for identifying fake news could have been learned or primed by the text and podcast treatments. Characteristics (1,2) are intended to positively signal truthfulness of a particular claim, while (3,4) negatively signal truthfulness. We test this by randomizing whether these features are associated with a given claim and then test whether treated respondents are more more likely to believe a claim when characteristics (1) and (2) are present and less likely to believe a claim when characteristics (3) and (4) are present. We combine these four measures using an ICW index. We expect that treated participants are likely to be more responsive to these signals than control, such that the interaction between treatment and the conjoint treatment is larger.

For H7, we measure effects on behavior relating to verifying information using questions asking: (1) how important they think fact-checking is; (2) how often they fact check information; (3) when they fact check, whether they use fact-checkers relative to other less reliable sources; (4)

whether they state that lack of knowledge about how and where to check information inhibits the extent of their fact-checking; and (5) whether they shared misinformation corrections with their friends and family. We combine these five measures using an ICW index.

The effects on these behavioral outcomes in H6 and H7 depend on how participants adjust to increased perceptions of misinformation, altered beliefs about the topics that were fact-checked, and/or empowerment to detect whether a piece of content constitutes misinformation.

3.5 Secondary treatment effects

We also examine potential secondary effects that the treatment may elicit. The posts that are fact-checked in the text messages and podcasts are topically broad. These fact-checks can be roughly divided into the following categories: (1) stoking anti-government or racial/nationalist sentiments from various important figures and politicians; (2) general conspiracy theories or fear-based misinformation; and (3) misinformation pertaining specifically to COVID-19 or vaccine hesitancy. The content of these podcasts could then influence related beliefs in several domains.

First, misinformation stemming from viral posts in categories (1) and (2) may promote political polarization and populist attitudes. We therefore hypothesize secondary treatment effects that temper such polarization:

H8 : Access to fact-checking content improves participants' perceptions of government performance and capacity and reduces support for populism.

We adapt questions on polarization and populism from various sources comprising: (1) perceptions of government performance, overall and with respect to COVID-19; (2) perceptions about government capacity (i.e. government's ability to carry out roads and electricity projects, conditional on its desire to do so); (3) beliefs about whether the government only serves elite interests; (4) whether the respondent intends to vote for the national incumbent party; and (5) whether the respondent feels close to the national incumbent party. We combine these outcomes using an ICW index.

Second, misinformation stemming from category (3) may discourage preventative behaviors while heightening fears around vaccination. We therefore test whether:

H9 : Access to fact-checking content increases participants' knowledge and beliefs in the severity of COVID-19 and their willingness to take preventative measures.

We measure this using questions relating to (1) self-reported preventative behavior in the week prior to enumeration; (2) beliefs in whether COVID-19 is a hoax and whether lockdowns are justified; and (3) trust in, and intentions to receive, a COVID-19 vaccine when available. We again combine these outcomes using an ICW index.

References

Brashier, Nadia M, Gordon Pennycook, Adam J Berinsky and David G Rand. 2021. "Timing matters when correcting fake news." *Proceedings of the National Academy of Sciences* 118(5).

Gaines, Brian J, James H Kuklinski, Paul J Quirk, Buddy Peyton and Jay Verkuilen. 2007. "Same facts, different interpretations: Partisan motivation and opinion on Iraq." *The Journal of Politics* 69(4):957–974.

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):15536–15545.

Jerit, Jennifer and Yangzi Zhao. 2020. "Political Misinformation." *Annual Review of Political Science* 23(1):77–94.
URL: *https://doi.org/10.1146/annurev-polisci-050718-032814*

Martel, C, G Pennycook and DG Rand. 2020. "Reliance on emotion promotes belief in fake news." *Cognitive Research: Principles and Implications* 5(47).

Nyhan, Brendan. 2020. "Facts and Myths about Misperceptions." *Journal of Economic Perspectives* 34(3):220–36.
URL: *https://www.aeaweb.org/articles?id=10.1257/jep.34.3.220*

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* pp. 1–6.

The Economist. 2019. "How WhatsApp is used and misused in Africa.".
URL: *https://www.economist.com/middle-east-and-africa/2019/07/18/how-whatsapp-is-used-and-misused-in-africa*

Van Bavel, Jay J and Andrea Pereira. 2018. "The partisan brain: An identity-based model of political belief." *Trends in cognitive sciences* 22(3):213–224.