Online Appendix

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

Table of Contents

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 three attention-checking questions within 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 WhatsApp numbers were also prevented from restarting the baseline survey.

A.2 Randomization

We enrolled participants in batches, once every two weeks. Individuals within each batch were blocked-randomized by demographics, social media consumption, trust towards different news sources, and knowledge about misinformation. Figure 3 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 Quiz administration

Participants were randomly assigned to take either *fact-check* quizzes (which served as incentivization to consume fact-checks) or *placebo* quizzes (which were meant to ensure similar levels of study engagement). Both quiz types were administered by the research team, and participants were asked six questions once every four weeks and informed that the quizzes were entirely voluntary. If they decided to take the quiz, they would earn R10 (0.62 USD) and would earn a further R10 if they answered at least four out of the six questions correctly. *Fact-check* quizzes covered information provided as part of the fact-checks over the past month, while *placebo* quizzes covered pop culture questions. Regardless of quiz type, participants were informed *how many* questions they answered correctly, but they were not told *which* questions they answered correctly. We did not provide answers to mitigate the risk of the quiz informing participants directly. Although the intervention did not forcibly inform participants of what is and is not true, it provided easy to use tools for participants to do so if they wanted to.

A.4 Financial compensation

We provided small financial compensation (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 smaller set of outcomes, respondents were provided R30 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 of the four quiz interim quizzes during the study (excluding midline and endline survey quizzes) during a given batch's study period, and the share of those completing each quiz who answered a majority of the questions correctly (and hence received high incentive payments).

A.5 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, regarding the intervention itself, our expectation was that each treatment arm would positively affect 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 potentially 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.

Second, regarding participation and consent, we solicited informed consent from all participants in the study and did not use deception 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.4), 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 leave 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, regarding the broader impact of the study, we expected that the limited sample size 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.6 Outcome measurement

All our main outcomes are inverse covariance weighted (ICW) indexes (see Anderson 2008). Each such outcome aggregates families of individual survey items, 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 2.

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. Due to capturing similar concepts, we merged hypotheses H2 and H3 in our pre-analysis plan into a combined H3 in the paper; Figures E3a and E4b report the results separately. Due to this merge and the order that results are presented in the paper, H2, H4, H5, H6, and H7 in the paper correspond to H5, H6, H4, H7, and H9 in our pre-analysis plan. We note below some deviations from our pre-specified measurement strategy; these changes were designed to focus attention on theoretically-relevant outcomes.

First, for exposure to the intervention (H1), we examine podcast take-up and knowledge of the content of the podcast separately to distinguish self-reported attention from internalization; we excluded a pre-specified index item about the frequency with which participants report being alerted to fake news across all social media platforms because we viewed this as a more general test of a distinct mechanism proposed in the literature on accuracy primes (e.g. Pennycook et al. 2021) rather than being a direct measure of exposure to our specific intervention; we find limited support for it (see Figure E1). We further examined *future* take-up as a separate indicator of treatment take-up once the small financial incentives to participate in the study had been removed, but this did not alter our pre-specified approach.

Second, for perceptions of misinformation and trust in social media (H3), the trust in social media component focuses on Facebook, Instagram, and Twitter. Because we merged perceptions of the extent of misinformation on social media, our index also included questions asking what source is trusted most for information and how much of the information received from social media platforms is likely to be true. We exclude all questions relating to WhatsApp because the fact-checking intervention was delivered via WhatsApp and hence results are difficult to interpret. Figure E4b shows that trust in information from close ties, again excluding information sent by these ties from WhatsApp, modestly decreases.

Third, for discernment (H4), our outcomes relating to conspiracy theories were not pre-registered due to their greater detachment from our treatments, but provide a valuable check on citizen evaluations of claims that could be the subject of misinformation. Additionally, we pre-registered the use of a conjoint experiment for the discernment outcomes. In its intended implementation, we sought to measure source credibility as a mechanism for discernment: respondents were meant to be randomly assigned to a slightly different version of each claim which added information relating to sourcing of that claim—whether from the National Institute on Alcohol and Alcoholism (NIAA, for the true claim about alcohol and COVID-19), Facebook (for the false claim about matric marks), the WHO (for the true claim about COVID-19 transmission), or the Ministry of Finance (for the false claim about foreign restaurant workers). We exclude this analysis due to an implementation failure which led nearly half of the batches in our study to be sent only one version of the relevant claim. In addition, interpretation of the individual findings was ambiguous. This was particularly true where the source may have seemed credible to participants at its face (Ministry of Finance) but the claim was false (foreign restaurant workers)—leading to possibly conflicting effects depending on whether discernment was driven by source credibility or increased skepticism. Within two other items, adding credible institutions—the NIAA and WHO—as information sources for true claims weakly reduced discernment in one instance (COVID-19 transmission claim) and had no clear effect on the other (alcohol reduces ability to fight COVID-19 infections). These results could be because credible sources are independent from, or serve as substitutes, for true claims. However, our limited power to conduct this analysis due to the implementation failure, plus the ambiguity of the results, means we do not present results of this analysis in the manuscript.

Fourth, for consumption and sharing of social media content (H5), we again exclude WhatsApp for the same reason as for H3. We also examine the consumption and sharing of information separately to examine effects on both important outcomes.

Fifth, for engaging in fact-checking (H6), 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 **??** 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 6a and Table F6). We exclude the pre-specified variable on whether respondents 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 (H7), 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 8b); (ii) we exclude two questions eliciting perceptions of government capacity (see Figure E6 for results) and two questions on populism-related beliefs (see Figure E7 for results), on the basis that these questions were worded to capture beliefs about how government *ought* to behave rather than concrete government appraisals; and (iii) we add willingness to vote regional incumbents to the index alongside our pre-specified measure of willingness to vote for the national government since the intervention could shape updating about either level of government.

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.

Second, results from questions that test participants' capacity to discern true from false news and their ability to identify conspiracy theories require knowledge of correct answers. 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.

Third, 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.

Finally, 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 E8 show that treated participants were indeed more likely to submit fact-check requests. Moreover, the greater effectiveness of the text message version of 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 *long* podcast

The *long* version of the “What’s Crap on WhatsApp?” podcast is available online through https://www.whatscrap.africa/.

B.3 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.4 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.

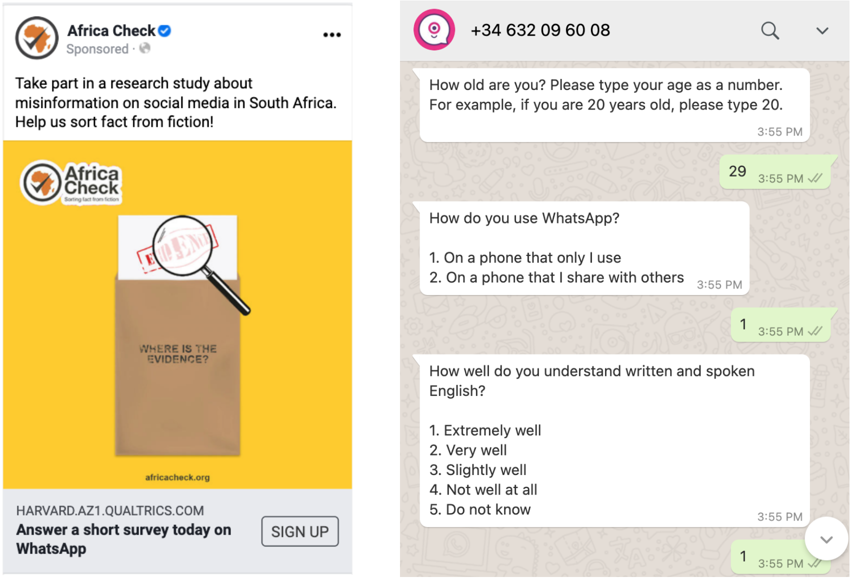
Overall, we find evidence consistent with the inclusion of a *Prime* message when encouraging participants to internalize their assigned treatments—particularly for the incentivized *Text* and *Empathetic* podcasts. These originally identified effects are then amplified by a *Prime* message which repeatedly 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.5 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



Africa Check
Sponsored ·

Take part in a research study about misinformation on social media in South Africa. Help us sort fact from fiction!

HARVARD.AZ1.QUALTRICS.COM
Answer a short survey today on WhatsApp

+34 632 09 60 08

How old are you? Please type your age as a number. For example, if you are 20 years old, please type 20. 3:55 PM

29 3:55 PM

How do you use WhatsApp?

1. On a phone that only I use
2. On a phone that I share with others 3:55 PM

1 3:55 PM

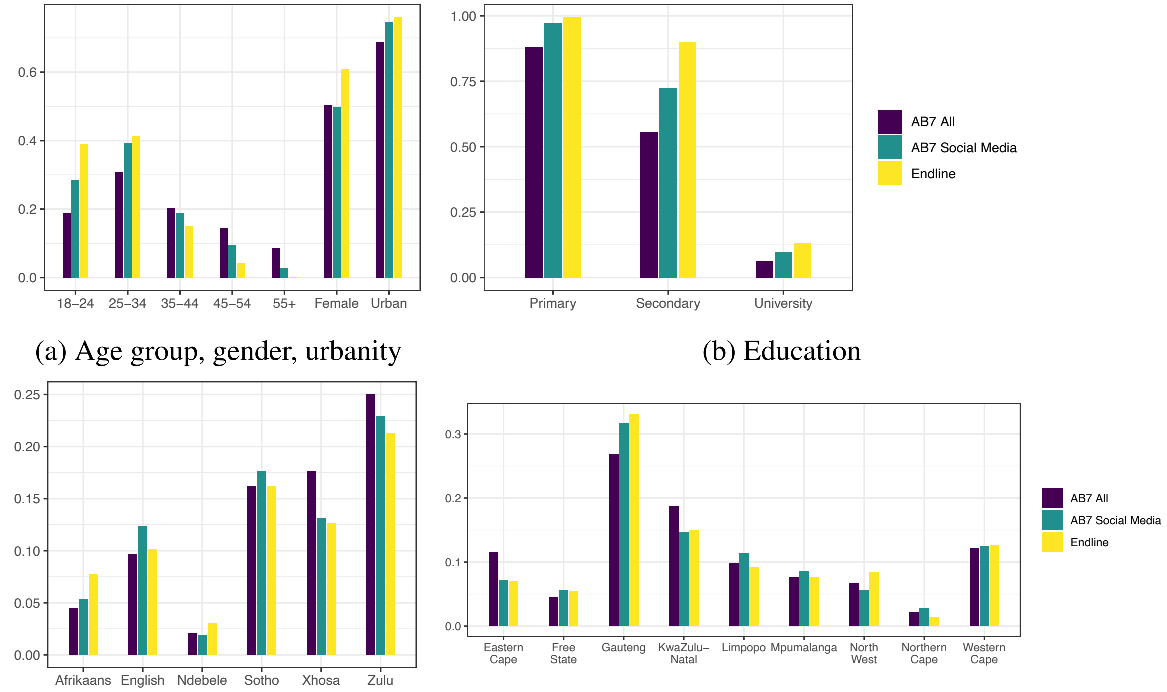
How well do you understand written and spoken English?

1. Extremely well
2. Very well
3. Slightly well
4. Not well at all
5. Do not know 3:55 PM

1 3:55 PM

(a) Recruitment Facebook ad (b) Survey through WhatsApp chatbot

Figure C1: Recruitment and surveying



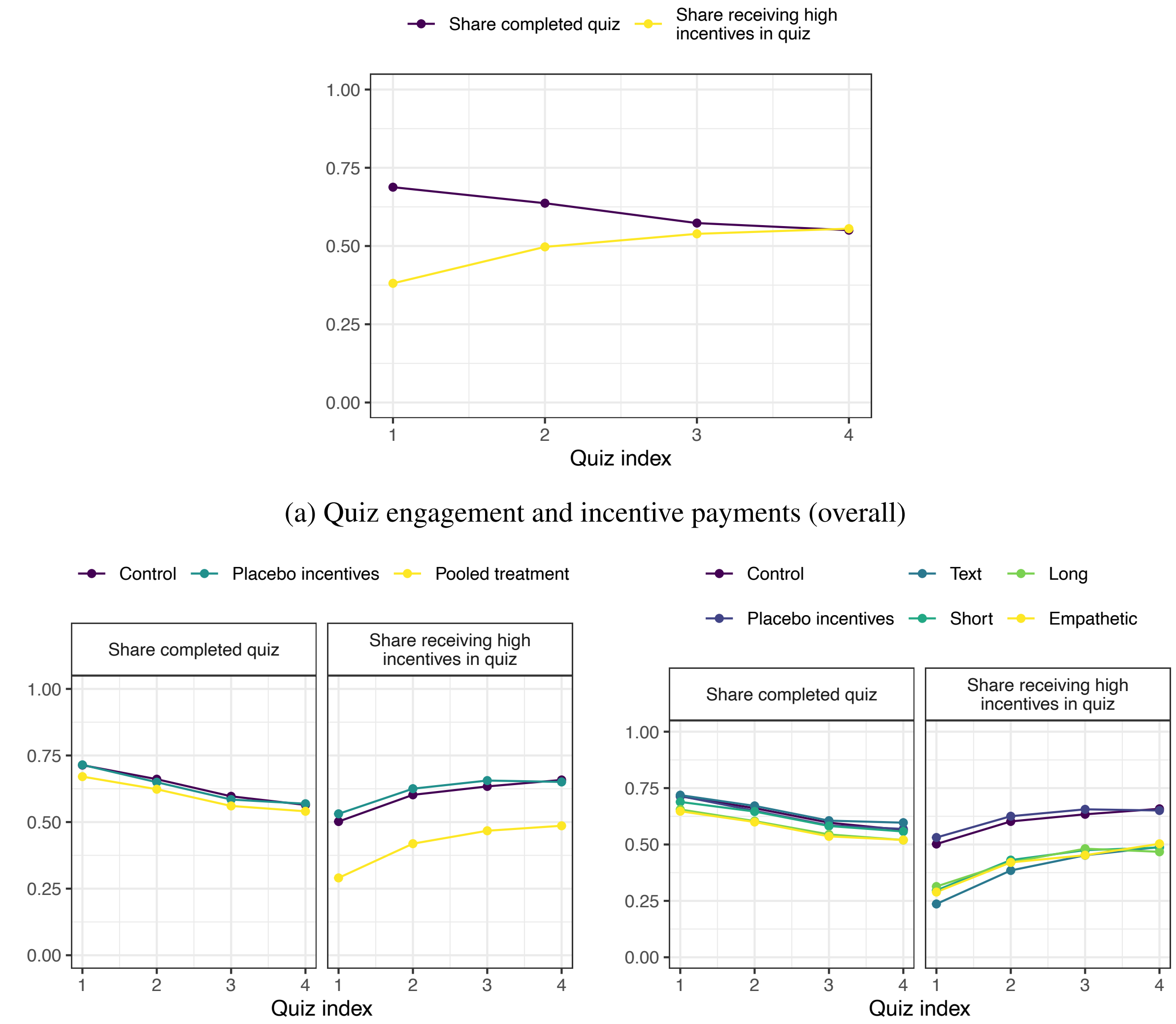
(a) Age group, gender, urbanity

(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 the average share of participants answering more than 50% of questions correctly (and hence receiving a larger incentive payment for completing the quiz), through each of the four study quizzes (fact-check or placebo) participants were sent between baseline and endline.

C.2 Balance and attrition

Table C1: Attrition

	Attrition	
	(1)	(2)
A. Pooled estimation		
Placebo incentives	0.023	0.021
	(0.017)	(0.017)
	[0.172]	[0.208]
Pooled treatment	-0.015	-0.017
	(0.012)	(0.012)
	[0.212]	[0.136]
B. Disaggregated estimation		
Placebo incentives	0.023	0.021
	(0.017)	(0.017)
	[0.171]	[0.207]
Text information	0.022	0.027
	(0.021)	(0.021)
	[0.301]	[0.189]
Short podcast	0.002	0.004
	(0.016)	(0.015)
	[0.909]	[0.803]
Long podcast	0.021	0.021
	(0.015)	(0.015)
	[0.166]	[0.168]
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

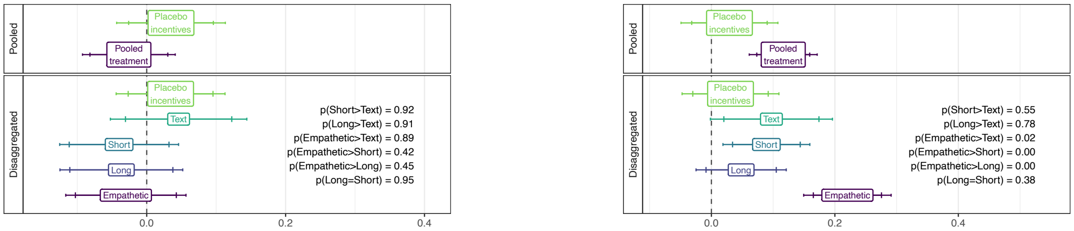
Notes: See Table 2 for variable definitions. All specifications are estimated using OLS, and adjust for randomization block fixed effects; even-indexed columns further include LASSO-selected controls. Heteroskedasticity-robust standard errors in parentheses, while p-values (adjusted for pre-registered direction when relevant) are in square brackets.

Table C2: Balance on pre-treatment outcomes

Variable	$p(\tau_{pooled} = 0)$	$p(\tau_{disagg.} = 0)$
A. Socio-demographic		
Gender: Female	[0.990]	[0.755]
Education: Primary	[0.367]	[0.017]
Education: Secondary	[0.855]	[0.757]
Education: University	[0.788]	[0.744]
Province: Eastern Cape	[0.372]	[0.693]
Province: Free State	[0.591]	[0.894]
Province: Gauteng	[0.871]	[0.995]
Province: KwaZulu-Natal	[0.828]	[0.409]
Province: Limpopo	[0.953]	[0.410]
Province: Mpumalanga	[0.528]	[0.129]
Province: Northern Cape	[0.096]	[0.386]
Province: North West	[0.204]	[0.551]
Province: Western Cape	[0.498]	[0.884]
Locality: Urban	[0.554]	[0.292]
Locality: Peri-urban	[0.569]	[0.908]
Locality: Rural	[0.551]	[0.800]
Age: 18-24	[0.786]	[0.650]
Age: 25-34	[0.180]	[0.481]
Age: 35-44	[0.530]	[0.769]
Age: 45-54	[0.164]	[0.061]
Age: 55+	[0.374]	[0.840]
B. Baseline survey responses		
Verify challenge	[0.411]	[0.784]
Consume news from close friends	[0.792]	[0.914]
Consume social media	[0.195]	[0.430]
Consume traditional media	[0.237]	[0.367]
Consume WhatsApp	[0.415]	[0.846]
COVID-19 beliefs and behavior	[0.153]	[0.456]
Non-WCW podcast take-up	[0.618]	[0.476]
Misinformation harmful	[0.893]	[0.511]
Podcast take-up	[0.885]	[0.910]
Sharing	[0.966]	[0.716]
Trust close friends	[0.961]	[0.561]
Trust organizations	[0.990]	[0.873]
Trust social media	[0.481]	[0.749]
Trust traditional media	[0.841]	[0.924]
Trust WhatsApp	[0.556]	[0.904]
Active verification	[0.711]	[0.216]
Verification knowledge	[0.163]	[0.260]

Notes: See Table 2 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.

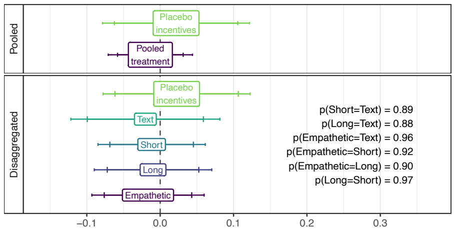
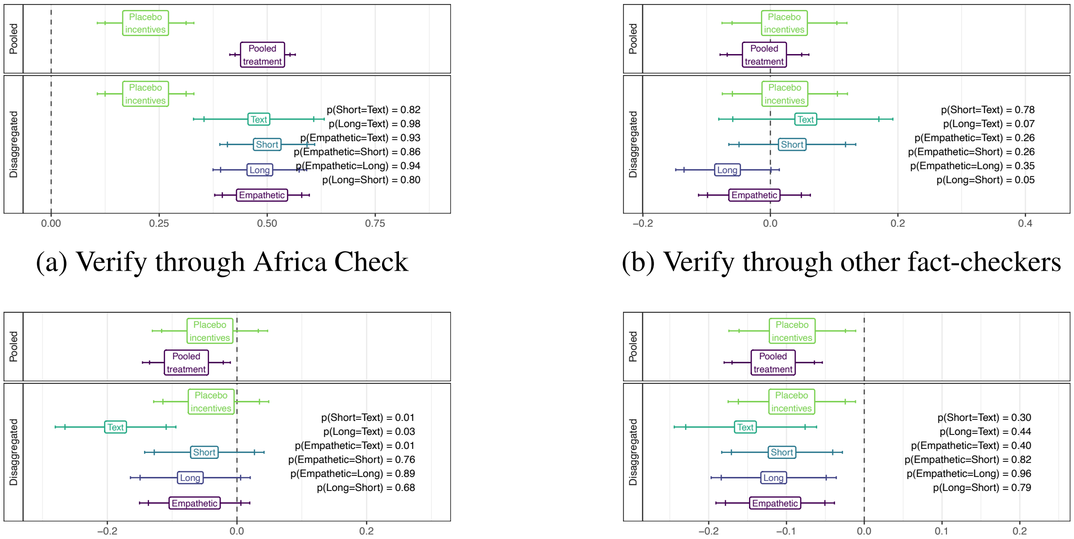


Figure D2: Treatment effects on difficulty of fact-checking

Notes: All outcomes are standardized inverse covariance-weighted indexes: 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 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.

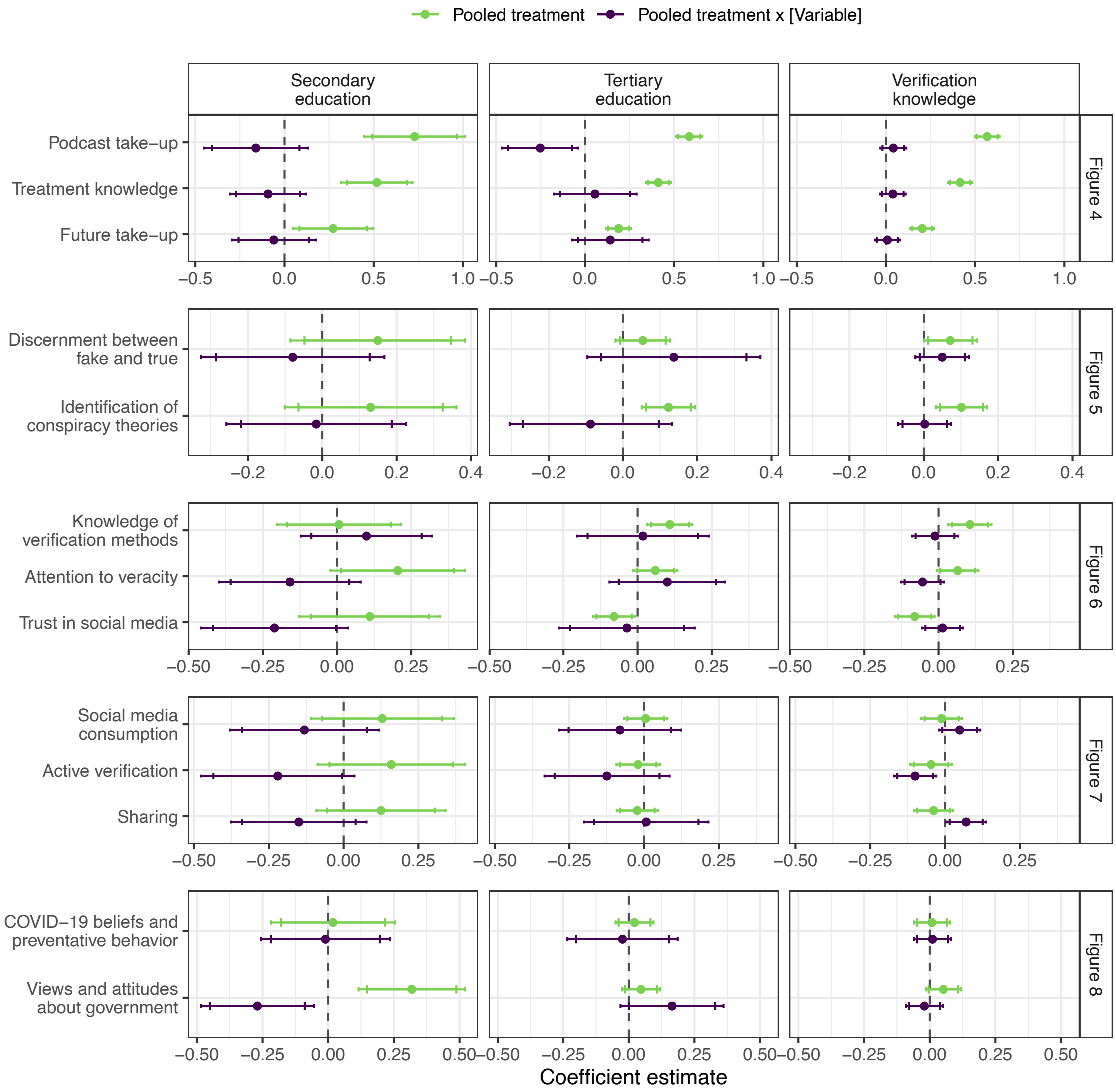


Figure D4: Heterogeneous treatment effects

Notes: All outcomes are standardized inverse covariance-weighted indexes corresponding to outcomes in the main manuscript (results split by which figure in the manuscript they correspond to). For a given row (outcome), the top coefficient represents the main effect of the pooled treatment while the bottom coefficient represents the coefficient on the interaction of the pooled treatment and a given pre-treatment covariate (which varies across the columns). 90% and 95% confidence intervals plotted.

E Figures referenced in supplementary materials and PAP

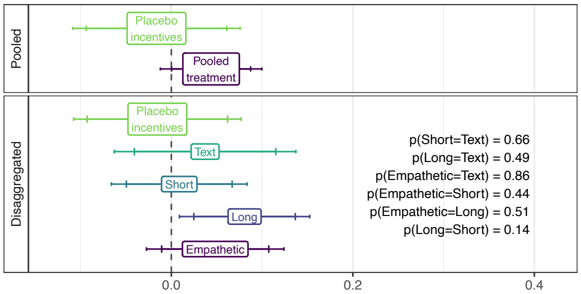


Figure E1: Being alerted about fake news

Notes: Outcome is standardized: How often participant is alerted about fake news. All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation (1); p-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals.

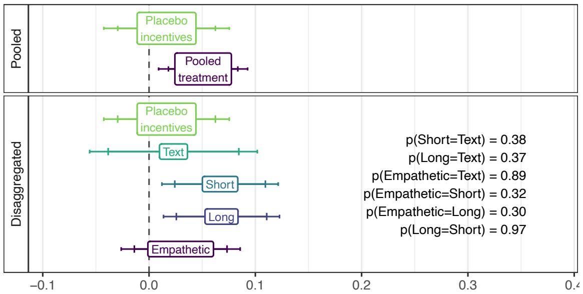
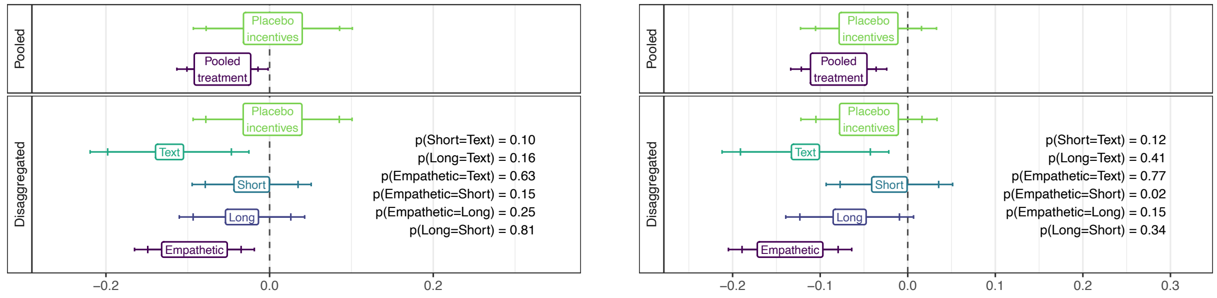


Figure E2: Alerting others about fake news

Notes: Outcome is standardized: How often participant reports alerting others about misinformation. All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation (1); p-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals.

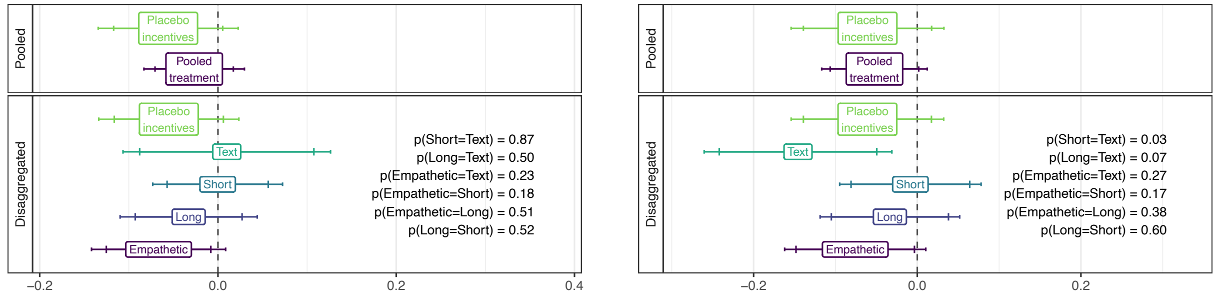


(a) Perceived truthfulness of social media content

(b) Trust in social media content

Figure E3: Disaggregating index on social media trust

Notes: All outcomes are standardized inverse covariance-weighted indexes: (a): believes information from social media likely to be true; (b) trusts information on social media, and thinks information on social media is most trustworthy. All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation (1); p-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals.

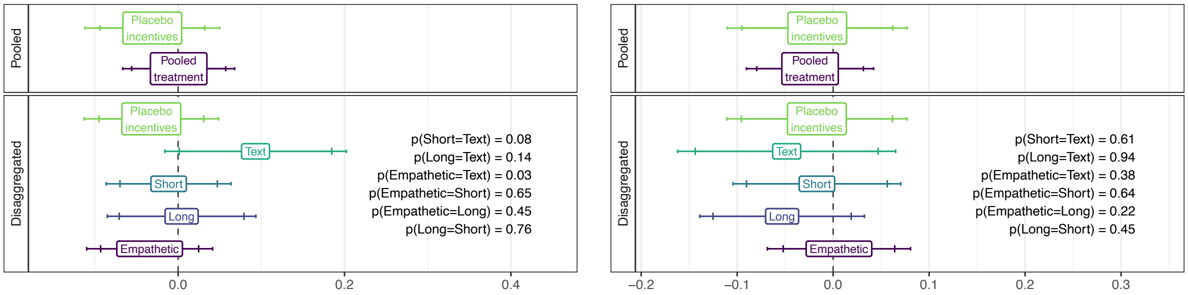


(a) Trust in traditional media

(b) Trust information sent by close ties

Figure E4: 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. All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation (1); p-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals.



(a) Traditional media consumption

(b) Consumption of news from close ties

Figure E5: 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. All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation (1); *p*-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals.

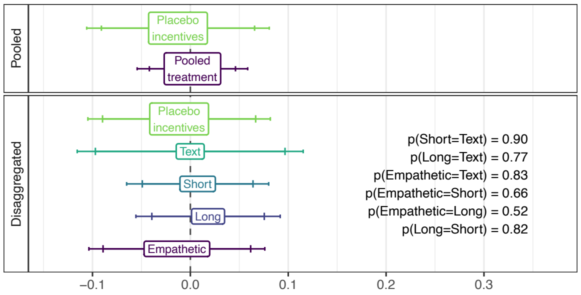


Figure E6: Perceptions of government capacity

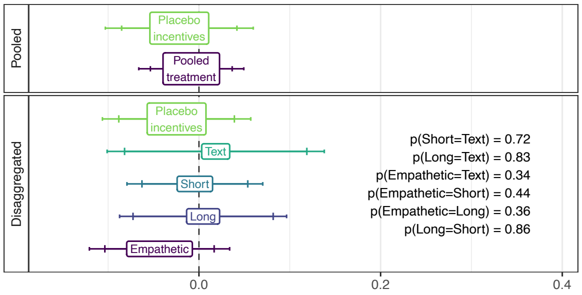


Figure E7: Populist attitudes

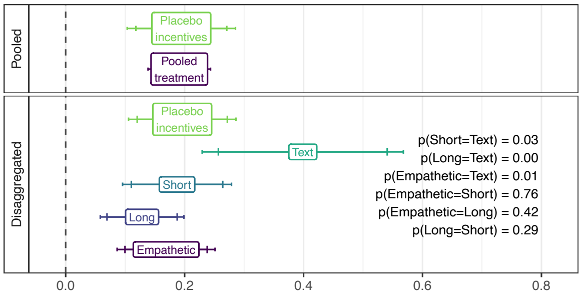


Figure E8: Fact-check requests

Notes: **Fig E6:** Outcome is standardized inverse covariance-weighted index comprising perception of government capacity to provide roads; perception of government capacity to supply electricity. **Fig E7:** Outcome is standardized inverse covariance-weighted index comprising perception of policies benefit elites; perception that ordinary people have no influence over policy. **Fig E8:** Outcome is a standardized indicator for participant submitting a fact-check request to Africa Check. All outcomes are standardized ICW indexes (see items in Table 2). Top panels within each subfigure provide pooled estimates of treatment effects; bottom panels provide estimates with disaggregated treatment variants. Estimated using Equation (1); *p*-values are from pre-registered tests of differences between treatment variants indicated in bottom panels, while the interior and exterior bars represent 90% and 95% confidence intervals.

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	0.415	0.426	0.017	0.029	0.247	0.253
	(0.054)	(0.054)	(0.059)	(0.059)	(0.025)	(0.024)
	[0.000]	[0.000]	[0.387]	[0.309]	[0.000]	[0.000]
Pooled podcast	0.651	0.650	0.131	0.122	0.360	0.362
	(0.036)	(0.035)	(0.041)	(0.041)	(0.015)	(0.015)
	[0.000]	[0.000]	[0.000]	[0.001]	[0.000]	[0.000]
B. Disaggregated estimation						
Placebo incentives	0.321	0.329	0.019	0.029	0.188	0.192
	(0.050)	(0.049)	(0.055)	(0.055)	(0.023)	(0.022)
	[0.000]	[0.000]	[0.362]	[0.296]	[0.000]	[0.000]
Text information	0.020	0.010	0.089	0.078	0.014	0.018
	(0.060)	(0.060)	(0.072)	(0.071)	(0.024)	(0.024)
	[0.742]	[0.866]	[0.220]	[0.270]	[0.281]	[0.228]
Short podcast	0.648	0.643	0.160	0.155	0.349	0.349
	(0.047)	(0.047)	(0.052)	(0.052)	(0.021)	(0.021)
	[0.000]	[0.000]	[0.001]	[0.001]	[0.000]	[0.000]
Long podcast	0.646	0.649	0.119	0.116	0.360	0.363
	(0.048)	(0.048)	(0.054)	(0.054)	(0.021)	(0.021)
	[0.000]	[0.000]	[0.013]	[0.016]	[0.000]	[0.000]
Empathetic podcast	0.663	0.661	0.114	0.100	0.374	0.376
	(0.048)	(0.048)	(0.054)	(0.052)	(0.021)	(0.021)
	[0.000]	[0.000]	[0.017]	[0.028]	[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	0.26	0.22	0.26	0.20	0.22
Observations	4543	4543	4543	4543	4543	4543

Notes: See Table 2 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 F2: Treatment knowledge

	ICW: Treatment knowledge		Fact-check quiz knowledge	
	(1)	(2)	(3)	(4)
A. Pooled estimation				
Placebo incentives	0.112	0.124	0.159	0.178
	(0.047)	(0.046)	(0.067)	(0.066)
	[0.008]	[0.004]	[0.008]	[0.003]
Pooled treatment	0.411	0.414	0.584	0.587
	(0.034)	(0.034)	(0.048)	(0.048)
	[0.000]	[0.000]	[0.000]	[0.000]
B. Disaggregated estimation				
Placebo incentives	0.113	0.126	0.160	0.179
	(0.047)	(0.046)	(0.067)	(0.066)
	[0.008]	[0.003]	[0.008]	[0.003]
Text information	0.336	0.339	0.476	0.481
	(0.064)	(0.062)	(0.091)	(0.087)
	[0.000]	[0.000]	[0.000]	[0.000]
Short podcast	0.388	0.386	0.551	0.547
	(0.046)	(0.045)	(0.065)	(0.064)
	[0.000]	[0.000]	[0.000]	[0.000]
Long podcast	0.373	0.384	0.530	0.545
	(0.048)	(0.047)	(0.068)	(0.066)
	[0.000]	[0.000]	[0.000]	[0.000]
Empathetic podcast	0.509	0.506	0.722	0.718
	(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	0.26	0.22	0.26
Observations	4543	4543	4543	4543

Notes: See Table 2 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 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	0.060	0.066	0.013	0.013	-0.003	0.000	0.030	0.032	0.049	0.050
	(0.050)	(0.049)	(0.021)	(0.021)	(0.019)	(0.019)	(0.023)	(0.023)	(0.023)	(0.023)
	[0.114]	[0.089]	[0.272]	[0.264]	[0.876]	[0.491]	[0.099]	[0.081]	[0.016]	[0.013]
Pooled treatment	0.204	0.208	0.139	0.140	0.052	0.054	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	0.060	0.067	0.013	0.013	0.003	0.000	0.030	0.032	0.049	0.050
	(0.050)	(0.049)	(0.021)	(0.021)	(0.019)	(0.019)	(0.023)	(0.023)	(0.023)	(0.023)
	[0.114]	[0.085]	[0.269]	[0.263]	[0.877]	[0.493]	[0.098]	[0.080]	[0.016]	[0.013]
Text information	0.213	0.237	0.019	0.025	0.065	0.072	0.081	0.092	0.083	0.090
	(0.057)	(0.055)	(0.026)	(0.026)	(0.021)	(0.021)	(0.028)	(0.027)	(0.028)	(0.028)
	[0.000]	[0.000]	[0.233]	[0.168]	[0.000]	[0.000]	[0.002]	[0.000]	[0.002]	[0.000]
Short podcast	0.234	0.240	0.149	0.151	0.061	0.063	0.094	0.097	0.103	0.104
	(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	0.171	0.171	0.168	0.166	0.039	0.041	0.068	0.067	0.085	0.083
	(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]	[0.010]	[0.007]	[0.000]	[0.000]	[0.000]	[0.000]
Empathetic podcast	0.202	0.200	0.156	0.155	0.048	0.050	0.083	0.081	0.093	0.091
	(0.044)	(0.042)	(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]	[0.001]	[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	0.47	0.47
R^2	0.09	0.14	0.11	0.14	0.08	0.11	0.08	0.13	0.08	0.11
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table 2 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 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	0.046	0.048	-0.020	-0.016	0.050	0.038	0.065	0.072	0.035	0.025
	(0.050)	(0.049)	(0.065)	(0.065)	(0.067)	(0.066)	(0.048)	(0.048)	(0.071)	(0.070)
	[0.179]	[0.165]	[0.758]	[0.811]	[0.226]	[0.283]	[0.088]	[0.066]	[0.311]	[0.359]
Pooled treatment	0.058	0.063	-0.126	-0.117	0.176	0.178	0.049	0.052	0.062	0.060
	(0.035)	(0.034)	(0.046)	(0.046)	(0.048)	(0.047)	(0.034)	(0.034)	(0.049)	(0.048)
	[0.048]	[0.034]	[0.006]	[0.011]	[0.000]	[0.000]	[0.077]	[0.064]	[0.102]	[0.106]
B. Disaggregated estimation										
Placebo incentives	0.046	0.050	0.020	0.013	0.050	0.032	0.065	0.074	0.035	0.025
	(0.050)	(0.049)	(0.065)	(0.065)	(0.067)	(0.066)	(0.048)	(0.048)	(0.071)	(0.070)
	[0.177]	[0.152]	[0.755]	[0.842]	[0.226]	[0.314]	[0.088]	[0.062]	[0.310]	[0.359]
Text information	0.121	0.112	0.001	0.013	0.195	0.162	0.060	0.067	0.044	0.027
	(0.063)	(0.062)	(0.079)	(0.079)	(0.081)	(0.079)	(0.057)	(0.058)	(0.088)	(0.087)
	[0.028]	[0.034]	[0.988]	[0.436]	[0.008]	[0.021]	[0.146]	[0.123]	[0.308]	[0.380]
Short podcast	0.026	0.026	0.154	0.146	0.153	0.145	0.052	0.045	0.023	0.024
	(0.046)	(0.045)	(0.061)	(0.061)	(0.062)	(0.060)	(0.043)	(0.043)	(0.062)	(0.061)
	[0.284]	[0.278]	[0.012]	[0.017]	[0.007]	[0.008]	[0.113]	[0.144]	[0.358]	[0.347]
Long podcast	0.018	0.001	0.160	0.153	0.087	0.091	0.047	0.056	0.020	0.015
	(0.046)	(0.046)	(0.063)	(0.063)	(0.064)	(0.063)	(0.046)	(0.045)	(0.066)	(0.065)
	[0.705]	[0.975]	[0.011]	[0.015]	[0.087]	[0.076]	[0.152]	[0.107]	[0.768]	[0.820]
Empathetic podcast	0.139	0.140	0.120	0.113	0.281	0.281	0.042	0.048	0.193	0.189
	(0.046)	(0.045)	(0.061)	(0.061)	(0.063)	(0.062)	(0.045)	(0.045)	(0.063)	(0.062)
	[0.001]	[0.000]	[0.049]	[0.065]	[0.000]	[0.000]	[0.179]	[0.142]	[0.001]	[0.001]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	-2.41	-2.41	2.77	2.77	-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	0.11	0.10	0.14
Observations	4543	4543	4145	4145	4145	4145	4145	4145	4145	4145

Notes: See Table 2 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 F5: Skepticism of conspiracy theories

	ICW: Conspiracy theories		AIDS invented (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	-0.023	-0.007	-0.095	-0.084	0.012	0.023	0.016	0.038	-0.010	0.009
	(0.050)	(0.049)	(0.070)	(0.068)	(0.070)	(0.068)	(0.067)	(0.066)	(0.069)	(0.068)
	[0.644]	[0.882]	[0.173]	[0.220]	[0.434]	[0.368]	[0.404]	[0.284]	[0.889]	[0.446]
Pooled treatment	0.105	0.111	0.071	0.081	0.092	0.102	0.179	0.185	0.112	0.117
	(0.035)	(0.034)	(0.048)	(0.048)	(0.048)	(0.047)	(0.048)	(0.047)	(0.047)	(0.047)
	[0.001]	[0.000]	[0.069]	[0.045]	[0.027]	[0.015]	[0.000]	[0.000]	[0.009]	[0.006]
B. Disaggregated estimation										
Placebo incentives	0.023	0.007	0.095	0.084	0.012	0.023	0.017	0.037	0.010	0.009
	(0.050)	(0.049)	(0.070)	(0.068)	(0.070)	(0.068)	(0.068)	(0.066)	(0.069)	(0.069)
	[0.646]	[0.883]	[0.173]	[0.218]	[0.434]	[0.370]	[0.402]	[0.288]	[0.889]	[0.450]
Text information	0.107	0.109	0.103	0.104	0.084	0.083	0.134	0.132	0.135	0.138
	(0.058)	(0.057)	(0.085)	(0.084)	(0.080)	(0.079)	(0.082)	(0.081)	(0.078)	(0.078)
	[0.033]	[0.028]	[0.111]	[0.107]	[0.147]	[0.145]	[0.051]	[0.052]	[0.043]	[0.040]
Short podcast	0.040	0.045	0.001	0.008	0.063	0.070	0.062	0.069	0.054	0.054
	(0.046)	(0.045)	(0.064)	(0.063)	(0.064)	(0.062)	(0.063)	(0.062)	(0.062)	(0.061)
	[0.196]	[0.157]	[0.496]	[0.450]	[0.162]	[0.129]	[0.161]	[0.133]	[0.195]	[0.188]
Long podcast	0.109	0.128	0.082	0.104	0.087	0.113	0.192	0.212	0.110	0.128
	(0.046)	(0.044)	(0.064)	(0.063)	(0.064)	(0.062)	(0.063)	(0.061)	(0.063)	(0.062)
	[0.008]	[0.002]	[0.098]	[0.050]	[0.086]	[0.035]	[0.001]	[0.000]	[0.040]	[0.019]
Empathetic podcast	0.167	0.166	0.119	0.122	0.131	0.126	0.307	0.302	0.165	0.160
	(0.045)	(0.043)	(0.063)	(0.062)	(0.063)	(0.061)	(0.060)	(0.059)	(0.061)	(0.060)
	[0.000]	[0.000]	[0.028]	[0.024]	[0.019]	[0.020]	[0.000]	[0.000]	[0.004]	[0.004]
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	0.11	0.08	0.15	0.08	0.12	0.07	0.11
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table 2 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 F6: Knowledge of verification methods (part 1)

	ICW: Verification knowledge		Avoid misinfo: Ask others (reversed)		Avoid misinfo: Seek reputable orgs		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)
A. Pooled estimation																
Placebo incentives	0.039	0.047	0.012	0.011	0.025	0.026	0.022	0.023	-0.021	-0.021	-0.011	-0.010	0.002	0.005	0.035	0.033
	(0.050)	(0.050)	(0.018)	(0.018)	(0.024)	(0.023)	(0.025)	(0.024)	(0.017)	(0.016)	(0.025)	(0.024)	(0.019)	(0.019)	(0.017)	(0.017)
	[0.216]	[0.173]	[0.245]	[0.271]	[0.149]	[0.134]	[0.185]	[0.174]	[0.198]	[0.203]	[0.654]	[0.671]	[0.460]	[0.396]	[0.022]	[0.030]
Pooled treatment	0.096	0.101	-0.020	-0.017	0.031	0.033	0.049	0.051	-0.013	-0.014	-0.014	-0.017	-0.001	-0.002	0.070	0.070
	(0.036)	(0.035)	(0.012)	(0.012)	(0.017)	(0.017)	(0.017)	(0.017)	(0.011)	(0.011)	(0.017)	(0.017)	(0.014)	(0.013)	(0.012)	(0.011)
	[0.003]	[0.002]	[0.099]	[0.149]	[0.031]	[0.024]	[0.002]	[0.001]	[0.235]	[0.197]	[0.396]	[0.312]	[0.955]	[0.857]	[0.000]	[0.000]
B. Disaggregated estimation																
Placebo incentives	0.039	0.047	0.012	0.011	0.025	0.026	0.022	0.023	0.021	0.021	0.011	0.010	0.002	0.005	0.035	0.033
	(0.050)	(0.050)	(0.018)	(0.018)	(0.024)	(0.023)	(0.025)	(0.024)	(0.017)	(0.016)	(0.025)	(0.024)	(0.019)	(0.019)	(0.017)	(0.017)
	[0.214]	[0.171]	[0.246]	[0.270]	[0.149]	[0.134]	[0.186]	[0.176]	[0.199]	[0.202]	[0.654]	[0.674]	[0.457]	[0.405]	[0.022]	[0.030]
Text information	0.167	0.176	0.011	0.011	0.036	0.038	0.071	0.072	0.031	0.033	0.009	0.012	0.011	0.011	0.039	0.036
	(0.064)	(0.064)	(0.022)	(0.022)	(0.030)	(0.030)	(0.031)	(0.030)	(0.020)	(0.020)	(0.030)	(0.030)	(0.024)	(0.024)	(0.021)	(0.021)
	[0.005]	[0.003]	[0.316]	[0.309]	[0.120]	[0.102]	[0.010]	[0.008]	[0.132]	[0.104]	[0.762]	[0.679]	[0.325]	[0.317]	[0.033]	[0.042]
Short podcast	0.124	0.125	0.003	0.002	0.010	0.008	0.034	0.034	0.001	0.001	0.006	0.007	0.010	0.008	0.076	0.074
	(0.048)	(0.047)	(0.016)	(0.016)	(0.022)	(0.022)	(0.023)	(0.022)	(0.014)	(0.014)	(0.022)	(0.022)	(0.018)	(0.018)	(0.016)	(0.016)
	[0.005]	[0.004]	[0.872]	[0.894]	[0.320]	[0.354]	[0.069]	[0.064]	[0.469]	[0.481]	[0.804]	[0.769]	[0.597]	[0.669]	[0.000]	[0.000]
Long podcast	0.022	0.034	0.032	0.030	0.035	0.040	0.056	0.062	0.012	0.014	0.018	0.023	0.018	0.022	0.063	0.066
	(0.048)	(0.048)	(0.015)	(0.015)	(0.023)	(0.022)	(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]	[0.050]	[0.062]	[0.036]	[0.007]	[0.003]	[0.415]	[0.333]	[0.438]	[0.314]	[0.325]	[0.225]	[0.000]	[0.000]
Empathetic podcast	0.110	0.111	0.039	0.034	0.048	0.049	0.046	0.047	0.021	0.022	0.023	0.024	0.020	0.017	0.087	0.086
	(0.049)	(0.049)	(0.015)	(0.015)	(0.023)	(0.022)	(0.023)	(0.023)	(0.015)	(0.015)	(0.022)	(0.022)	(0.018)	(0.017)	(0.017)	(0.017)
	[0.012]	[0.012]	[0.010]	[0.024]	[0.017]	[0.014]	[0.021]	[0.018]	[0.164]	[0.145]	[0.311]	[0.279]	[0.122]	[0.162]	[0.000]	[0.000]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	0.14	0.14	0.34	0.34	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	0.49	0.49	0.31	0.31	0.48	0.48	0.38	0.38	0.31	0.31
R^2	0.09	0.10	0.06	0.09	0.06	0.09	0.07	0.11	0.07	0.09	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

Notes: See Table 2 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 F6: 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 FC		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	0.047	0.006	0.009	0.000	0.001	0.008	0.010	-0.017	-0.015	0.017	0.021	0.026	0.028	-0.023	-0.023
	(0.050)	(0.050)	(0.019)	(0.019)	(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]	[0.172]	[0.370]	[0.318]	[0.991]	[0.480]	[0.283]	[0.243]	[0.321]	[0.376]	[0.190]	[0.133]	[0.144]	[0.129]	[0.338]	[0.336]
Pooled treatment	0.096	0.101	-0.014	-0.014	0.027	0.025	0.005	0.005	0.007	0.005	0.050	0.050	0.053	0.057	-0.012	-0.009
	(0.036)	(0.035)	(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]	[0.002]	[0.315]	[0.310]	[0.045]	[0.053]	[0.311]	[0.307]	[0.275]	[0.327]	[0.000]	[0.000]	[0.001]	[0.000]	[0.495]	[0.590]
B. Disaggregated estimation																
Placebo incentives	0.039	0.048	0.006	0.008	0.000	0.001	0.008	0.010	0.017	0.015	0.017	0.020	0.026	0.028	0.023	0.021
	(0.050)	(0.050)	(0.019)	(0.019)	(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]	[0.167]	[0.370]	[0.330]	[0.991]	[0.475]	[0.281]	[0.242]	[0.319]	[0.374]	[0.189]	[0.152]	[0.144]	[0.123]	[0.342]	[0.380]
Text information	0.167	0.176	0.007	0.006	0.056	0.058	0.010	0.007	0.026	0.028	0.075	0.077	0.087	0.091	0.016	0.011
	(0.064)	(0.064)	(0.024)	(0.023)	(0.027)	(0.027)	(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]	[0.805]	[0.020]	[0.014]	[0.294]	[0.343]	[0.085]	[0.069]	[0.002]	[0.001]	[0.002]	[0.001]	[0.592]	[0.715]
Short podcast	0.124	0.124	0.011	0.012	0.002	0.002	0.004	0.004	0.010	0.009	0.059	0.057	0.053	0.053	0.001	0.002
	(0.048)	(0.047)	(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)	(0.023)
	[0.005]	[0.004]	[0.552]	[0.508]	[0.469]	[0.455]	[0.397]	[0.382]	[0.244]	[0.282]	[0.000]	[0.001]	[0.011]	[0.011]	[0.487]	[0.469]
Long podcast	0.022	0.034	0.019	0.019	0.017	0.013	0.008	0.007	0.007	0.004	0.027	0.028	0.046	0.051	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)	(0.023)
	[0.324]	[0.242]	[0.293]	[0.282]	[0.206]	[0.267]	[0.560]	[0.608]	[0.327]	[0.393]	[0.071]	[0.058]	[0.025]	[0.013]	[0.133]	[0.248]
Empathetic podcast	0.110	0.111	0.014	0.012	0.050	0.047	0.018	0.018	0.005	0.008	0.052	0.052	0.046	0.052	0.000	0.002
	(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]	[0.012]	[0.432]	[0.503]	[0.007]	[0.010]	[0.087]	[0.088]	[0.724]	[0.596]	[0.003]	[0.003]	[0.024]	[0.013]	[1.000]	[0.938]
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	0.13	0.08	0.10	0.07	0.10	0.07	0.11	0.08	0.11	0.11	0.13
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table 2 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 F7: Attention to veracity of social media content

	ICW: Attention to veracity		Avoid misinfo: Check source		How important to verify		How often think twice	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Pooled estimation								
Placebo incentives	0.039	0.044	0.025	0.025	0.000	0.009	-0.013	-0.002
	(0.050)	(0.049)	(0.024)	(0.024)	(0.061)	(0.060)	(0.053)	(0.052)
	[0.219]	[0.187]	[0.157]	[0.148]	[0.498]	[0.442]	[0.804]	[0.972]
Pooled treatment	0.054	0.058	0.032	0.033	0.035	0.036	-0.036	-0.030
	(0.035)	(0.034)	(0.017)	(0.017)	(0.043)	(0.042)	(0.037)	(0.036)
	[0.061]	[0.044]	[0.032]	[0.027]	[0.209]	[0.193]	[0.342]	[0.408]
B. Disaggregated estimation								
Placebo incentives	0.040	0.044	0.025	0.025	0.000	0.009	0.013	0.003
	(0.050)	(0.049)	(0.024)	(0.024)	(0.061)	(0.060)	(0.053)	(0.052)
	[0.215]	[0.184]	[0.154]	[0.148]	[0.497]	[0.437]	[0.806]	[0.957]
Text information	0.007	0.017	0.018	0.016	0.037	0.016	0.062	0.036
	(0.064)	(0.061)	(0.030)	(0.030)	(0.079)	(0.076)	(0.065)	(0.063)
	[0.455]	[0.389]	[0.276]	[0.293]	[0.638]	[0.832]	[0.336]	[0.566]
Short podcast	0.077	0.072	0.045	0.042	0.013	0.008	0.008	0.011
	(0.046)	(0.045)	(0.023)	(0.022)	(0.056)	(0.054)	(0.049)	(0.047)
	[0.046]	[0.055]	[0.024]	[0.029]	[0.410]	[0.442]	[0.863]	[0.823]
Long podcast	0.010	0.000	0.008	0.003	0.062	0.067	0.053	0.046
	(0.047)	(0.045)	(0.023)	(0.022)	(0.057)	(0.056)	(0.050)	(0.048)
	[0.829]	[0.496]	[0.739]	[0.894]	[0.140]	[0.114]	[0.286]	[0.339]
Empathetic podcast	0.117	0.121	0.064	0.066	0.065	0.061	0.034	0.030
	(0.046)	(0.045)	(0.023)	(0.023)	(0.057)	(0.055)	(0.049)	(0.048)
	[0.006]	[0.004]	[0.003]	[0.002]	[0.127]	[0.132]	[0.485]	[0.523]
Controls	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	0.00	0.00	0.38	0.38	4.04	4.04	3.86	3.86
Control SD	1.00	1.00	0.49	0.49	1.25	1.25	1.06	1.06
R^2	0.07	0.13	0.07	0.10	0.07	0.13	0.07	0.13
Observations	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table 2 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 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	-0.035	0.004	0.002	-0.023	-0.023	-0.012	-0.016
	(0.047)	(0.046)	(0.038)	(0.036)	(0.019)	(0.018)	(0.050)	(0.049)
	[0.232]	[0.229]	[0.909]	[0.955]	[0.111]	[0.108]	[0.402]	[0.376]
Pooled treatment	-0.087	-0.084	-0.049	-0.043	-0.035	-0.031	-0.047	-0.050
	(0.034)	(0.033)	(0.026)	(0.025)	(0.014)	(0.013)	(0.035)	(0.035)
	[0.005]	[0.006]	[0.029]	[0.041]	[0.005]	[0.011]	[0.092]	[0.073]
B. Disaggregated estimation								
Placebo incentives	0.035	0.035	0.004	0.004	0.023	0.023	0.013	0.016
	(0.047)	(0.046)	(0.038)	(0.036)	(0.019)	(0.018)	(0.050)	(0.050)
	[0.232]	[0.227]	[0.910]	[0.918]	[0.111]	[0.105]	[0.401]	[0.375]
Text information	0.152	0.140	0.102	0.090	0.055	0.050	0.064	0.056
	(0.058)	(0.057)	(0.044)	(0.043)	(0.022)	(0.022)	(0.062)	(0.061)
	[0.004]	[0.007]	[0.011]	[0.019]	[0.008]	[0.011]	[0.150]	[0.178]
Short podcast	0.022	0.022	0.024	0.015	0.010	0.007	0.006	0.013
	(0.044)	(0.043)	(0.034)	(0.032)	(0.018)	(0.018)	(0.046)	(0.045)
	[0.308]	[0.302]	[0.232]	[0.317]	[0.279]	[0.356]	[0.451]	[0.390]
Long podcast	0.067	0.067	0.024	0.024	0.033	0.030	0.028	0.039
	(0.045)	(0.044)	(0.035)	(0.034)	(0.018)	(0.017)	(0.047)	(0.047)
	[0.068]	[0.064]	[0.250]	[0.241]	[0.032]	[0.043]	[0.273]	[0.204]
Empathetic podcast	0.147	0.139	0.074	0.067	0.053	0.048	0.099	0.098
	(0.043)	(0.043)	(0.034)	(0.033)	(0.017)	(0.017)	(0.046)	(0.045)
	[0.000]	[0.000]	[0.014]	[0.020]	[0.000]	[0.002]	[0.016]	[0.016]
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	0.17	0.10	0.17	0.07	0.10	0.14	0.17
Observations	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table 2 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 F9: Social media consumption

	ICW: Consume social media		Get news from: Other social media	
	(1)	(2)	(3)	(4)
A. Pooled estimation				
Placebo incentives	-0.016	-0.022	-0.016	-0.011
	(0.049)	(0.048)	(0.024)	(0.024)
	[0.372]	[0.323]	[0.259]	[0.323]
Pooled treatment	-0.005	-0.008	-0.008	-0.004
	(0.034)	(0.034)	(0.017)	(0.017)
	[0.438]	[0.400]	[0.313]	[0.400]
B. Disaggregated estimation				
Placebo incentives	0.016	0.022	0.016	0.011
	(0.049)	(0.048)	(0.024)	(0.024)
	[0.372]	[0.323]	[0.259]	[0.324]
Text information	0.072	0.066	0.038	0.033
	(0.060)	(0.059)	(0.030)	(0.029)
	[0.117]	[0.132]	[0.105]	[0.135]
Short podcast	0.021	0.022	0.007	0.012
	(0.045)	(0.045)	(0.023)	(0.022)
	[0.635]	[0.625]	[0.744]	[0.603]
Long podcast	0.022	0.011	0.001	0.006
	(0.045)	(0.045)	(0.023)	(0.022)
	[0.623]	[0.807]	[0.954]	[0.791]
Empathetic podcast	0.030	0.036	0.021	0.018
	(0.045)	(0.044)	(0.022)	(0.022)
	[0.249]	[0.209]	[0.176]	[0.211]
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	0.14
Observations	4543	4543	4543	4543

Notes: See Table 2 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 F10: Active verification

	ICW: Active verification		How often verify	
	(1)	(2)	(3)	(4)
A. Pooled estimation				
Placebo incentives	-0.040	-0.032	-0.044	-0.035
	(0.048)	(0.048)	(0.054)	(0.053)
	[0.409]	[0.505]	[0.409]	[0.505]
Pooled treatment	-0.039	-0.036	-0.043	-0.040
	(0.034)	(0.034)	(0.038)	(0.037)
	[0.258]	[0.283]	[0.258]	[0.283]
B. Disaggregated estimation				
Placebo incentives	0.040	0.034	0.044	0.036
	(0.048)	(0.048)	(0.054)	(0.053)
	[0.407]	[0.481]	[0.407]	[0.502]
Text information	0.127	0.127	0.141	0.141
	(0.065)	(0.064)	(0.072)	(0.071)
	[0.050]	[0.045]	[0.050]	[0.046]
Short podcast	0.042	0.040	0.046	0.044
	(0.045)	(0.044)	(0.049)	(0.049)
	[0.348]	[0.369]	[0.348]	[0.363]
Long podcast	0.016	0.015	0.017	0.017
	(0.043)	(0.043)	(0.048)	(0.048)
	[0.360]	[0.363]	[0.360]	[0.359]
Empathetic podcast	0.049	0.042	0.054	0.048
	(0.046)	(0.045)	(0.051)	(0.050)
	[0.282]	[0.348]	[0.282]	[0.338]
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	4543	4543	4543	4543

Notes: See Table 2 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 F11: Sharing

	ICW: Sharing		How often share stories	
	(1)	(2)	(3)	(4)
A. Pooled estimation				
Placebo incentives	0.021	0.012	0.021	0.015
	(0.046)	(0.045)	(0.054)	(0.051)
	[0.656]	[0.796]	[0.697]	[0.767]
Pooled treatment	-0.029	-0.029	-0.035	-0.031
	(0.033)	(0.032)	(0.039)	(0.036)
	[0.193]	[0.184]	[0.183]	[0.194]
B. Disaggregated estimation				
Placebo incentives	0.021	0.014	0.021	0.015
	(0.046)	(0.045)	(0.054)	(0.051)
	[0.657]	[0.750]	[0.698]	[0.772]
Text information	0.103	0.093	0.121	0.108
	(0.057)	(0.054)	(0.065)	(0.061)
	[0.035]	[0.043]	[0.032]	[0.039]
Short podcast	0.020	0.025	0.023	0.025
	(0.044)	(0.042)	(0.051)	(0.048)
	[0.641]	[0.547]	[0.653]	[0.601]
Long podcast	0.003	0.007	0.004	0.009
	(0.044)	(0.043)	(0.051)	(0.049)
	[0.469]	[0.432]	[0.933]	[0.424]
Empathetic podcast	0.071	0.069	0.095	0.078
	(0.043)	(0.041)	(0.050)	(0.047)
	[0.048]	[0.049]	[0.029]	[0.049]
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	0.23	0.12	0.23
Observations	4543	4543	4543	4543

Notes: See Table 2 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 7c.

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 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	-0.031	-0.070	-0.081	-0.109	-0.096	0.168	0.182	0.068	0.079	-0.039	-0.027	-0.040	-0.025	-0.029	-0.018
	(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)
	[0.393]	[0.519]	[0.514]	[0.445]	[0.290]	[0.341]	[0.070]	[0.055]	[0.110]	[0.074]	[0.382]	[0.551]	[0.552]	[0.714]	[0.713]	[0.817]
Pooled treatment	0.003	0.006	-0.032	-0.040	-0.028	-0.020	0.048	0.051	0.083	0.093	-0.016	-0.007	0.028	0.034	0.043	0.047
	(0.034)	(0.033)	(0.076)	(0.076)	(0.071)	(0.070)	(0.080)	(0.080)	(0.039)	(0.039)	(0.032)	(0.032)	(0.049)	(0.048)	(0.055)	(0.055)
	[0.464]	[0.424]	[0.672]	[0.593]	[0.692]	[0.772]	[0.275]	[0.261]	[0.017]	[0.008]	[0.622]	[0.825]	[0.287]	[0.242]	[0.217]	[0.193]
B. Disaggregated estimation																
Placebo incentives	0.041	0.033	0.070	0.077	0.109	0.090	0.167	0.180	0.068	0.079	0.039	0.026	0.041	0.025	0.029	0.019
	(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]	[0.496]	[0.511]	[0.470]	[0.289]	[0.371]	[0.072]	[0.056]	[0.110]	[0.074]	[0.387]	[0.561]	[0.550]	[0.711]	[0.706]	[0.810]
Text information	0.142	0.161	0.052	0.032	0.263	0.266	0.272	0.284	0.092	0.103	0.062	0.054	0.050	0.077	0.123	0.133
	(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]	[0.003]	[0.344]	[0.404]	[0.017]	[0.014]	[0.017]	[0.013]	[0.086]	[0.062]	[0.275]	[0.339]	[0.277]	[0.175]	[0.093]	[0.074]
Short podcast	0.020	0.027	0.004	0.014	0.034	0.033	0.091	0.076	0.113	0.121	0.041	0.047	0.056	0.060	0.055	0.055
	(0.044)	(0.043)	(0.101)	(0.101)	(0.094)	(0.092)	(0.105)	(0.104)	(0.051)	(0.050)	(0.042)	(0.041)	(0.064)	(0.063)	(0.072)	(0.072)
	[0.327]	[0.269]	[0.965]	[0.893]	[0.719]	[0.720]	[0.193]	[0.231]	[0.013]	[0.008]	[0.161]	[0.127]	[0.192]	[0.171]	[0.224]	[0.220]
Long podcast	0.024	0.014	0.017	0.018	0.127	0.103	0.068	0.076	0.059	0.076	0.056	0.045	0.045	0.049	0.090	0.093
	(0.047)	(0.046)	(0.101)	(0.101)	(0.099)	(0.096)	(0.106)	(0.106)	(0.052)	(0.052)	(0.043)	(0.043)	(0.065)	(0.064)	(0.072)	(0.071)
	[0.605]	[0.764]	[0.866]	[0.856]	[0.197]	[0.287]	[0.262]	[0.237]	[0.128]	[0.069]	[0.195]	[0.292]	[0.243]	[0.224]	[0.105]	[0.096]
Empathetic podcast	0.051	0.050	0.114	0.116	0.056	0.059	0.119	0.106	0.073	0.078	0.013	0.008	0.030	0.029	0.054	0.046
	(0.045)	(0.044)	(0.101)	(0.100)	(0.095)	(0.093)	(0.109)	(0.108)	(0.052)	(0.051)	(0.042)	(0.042)	(0.064)	(0.063)	(0.073)	(0.072)
	[0.259]	[0.263]	[0.257]	[0.247]	[0.554]	[0.525]	[0.276]	[0.326]	[0.081]	[0.065]	[0.751]	[0.846]	[0.641]	[0.645]	[0.465]	[0.520]
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	0.14	0.15	0.16	0.10	0.14	0.14	0.16	0.08	0.11	0.09	0.11	0.07	0.11	0.06	0.09
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table 2 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 F13: Government attitudes

	ICW: Govt attitudes		General govt performance		Gov handled COVID well		How true: Info from pols		Trust most for info: Gov		Trust most for info: Pols		Trust: Info from politicians		Vote: National incumbent		Vote: Regional incumbent	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
A. Pooled estimation																		
Placebo incentives	0.094	0.099	0.079	0.081	0.026	0.033	-0.030	-0.026	0.009	0.012	0.019	0.018	-0.013	-0.016	0.033	0.033	0.060	0.060
	(0.050)	(0.047)	(0.059)	(0.057)	(0.061)	(0.060)	(0.048)	(0.047)	(0.023)	(0.022)	(0.017)	(0.017)	(0.059)	(0.056)	(0.020)	(0.020)	(0.021)	(0.021)
	[0.030]	[0.019]	[0.089]	[0.078]	[0.334]	[0.292]	[0.525]	[0.583]	[0.343]	[0.303]	[0.132]	[0.143]	[0.827]	[0.780]	[0.055]	[0.051]	[0.002]	[0.002]
Pooled treatment	0.060	0.062	0.051	0.046	0.027	0.036	-0.033	-0.031	0.021	0.023	0.020	0.020	-0.035	-0.033	0.007	0.005	0.020	0.021
	(0.035)	(0.033)	(0.042)	(0.040)	(0.043)	(0.042)	(0.033)	(0.033)	(0.016)	(0.016)	(0.012)	(0.011)	(0.041)	(0.040)	(0.014)	(0.014)	(0.014)	(0.014)
	[0.042]	[0.031]	[0.109]	[0.126]	[0.264]	[0.197]	[0.323]	[0.345]	[0.090]	[0.076]	[0.046]	[0.038]	[0.396]	[0.415]	[0.322]	[0.365]	[0.081]	[0.067]
B. Disaggregated estimation																		
Placebo incentives	0.094	0.098	0.079	0.082	0.027	0.035	0.030	0.025	0.009	0.012	0.019	0.019	0.013	0.014	0.032	0.032	0.059	0.059
	(0.050)	(0.047)	(0.059)	(0.057)	(0.061)	(0.060)	(0.048)	(0.047)	(0.023)	(0.022)	(0.017)	(0.017)	(0.059)	(0.056)	(0.020)	(0.020)	(0.021)	(0.021)
	[0.030]	[0.019]	[0.089]	[0.073]	[0.331]	[0.278]	[0.528]	[0.591]	[0.342]	[0.304]	[0.131]	[0.131]	[0.827]	[0.805]	[0.056]	[0.056]	[0.003]	[0.002]
Text information	0.074	0.090	0.033	0.039	0.062	0.036	0.037	0.053	0.047	0.051	0.000	0.003	0.009	0.005	0.041	0.045	0.055	0.061
	(0.061)	(0.058)	(0.069)	(0.068)	(0.074)	(0.073)	(0.057)	(0.056)	(0.030)	(0.029)	(0.020)	(0.020)	(0.076)	(0.073)	(0.026)	(0.025)	(0.026)	(0.026)
	[0.114]	[0.061]	[0.316]	[0.283]	[0.405]	[0.621]	[0.258]	[0.174]	[0.057]	[0.039]	[0.492]	[0.433]	[0.910]	[0.474]	[0.054]	[0.036]	[0.017]	[0.008]
Short podcast	0.120	0.115	0.095	0.089	0.118	0.119	0.015	0.015	0.032	0.030	0.026	0.027	0.017	0.012	0.020	0.016	0.032	0.030
	(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)	(0.019)	(0.019)
	[0.005]	[0.004]	[0.043]	[0.047]	[0.017]	[0.015]	[0.361]	[0.360]	[0.067]	[0.078]	[0.046]	[0.039]	[0.377]	[0.412]	[0.143]	[0.191]	[0.050]	[0.056]
Long podcast	0.038	0.036	0.032	0.021	0.013	0.004	0.098	0.102	0.000	0.003	0.020	0.020	0.062	0.068	0.017	0.012	0.034	0.032
	(0.048)	(0.046)	(0.056)	(0.054)	(0.057)	(0.056)	(0.045)	(0.044)	(0.021)	(0.021)	(0.016)	(0.016)	(0.056)	(0.055)	(0.019)	(0.019)	(0.020)	(0.019)
	[0.212]	[0.217]	[0.283]	[0.353]	[0.822]	[0.942]	[0.028]	[0.020]	[0.993]	[0.445]	[0.103]	[0.098]	[0.270]	[0.215]	[0.193]	[0.253]	[0.041]	[0.048]
Empathetic podcast	0.013	0.022	0.034	0.038	0.013	0.025	0.049	0.043	0.020	0.022	0.021	0.022	0.075	0.068	0.034	0.031	0.023	0.017
	(0.047)	(0.045)	(0.055)	(0.053)	(0.056)	(0.055)	(0.045)	(0.044)	(0.021)	(0.021)	(0.016)	(0.016)	(0.055)	(0.053)	(0.018)	(0.018)	(0.018)	(0.018)
	[0.392]	[0.316]	[0.269]	[0.239]	[0.407]	[0.323]	[0.276]	[0.333]	[0.169]	[0.153]	[0.085]	[0.075]	[0.173]	[0.201]	[0.063]	[0.079]	[0.219]	[0.338]
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	0.20	0.20	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	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	0.17	0.08	0.14	0.08	0.12
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: See Table 2 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 F14: Pooled estimation ICW outcomes (including LASSO-selected covariate coefficients)

Figure	4a	4b	4c	5a	5	6a	6b	6c	7a	7b	7c	8a	8b
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Placebo incentives	0.33 (0.05)	0.12 (0.05)	0.07 (0.05)	0.05 (0.05)	0.01 (0.05)	0.05 (0.05)	0.04 (0.05)	0.03 (0.05)	0.02 (0.05)	0.03 (0.05)	0.01 (0.04)	0.03 (0.05)	0.10 (0.05)
Pooled treatment	0.57 (0.03)	0.41 (0.03)	0.21 (0.03)	0.06 (0.03)	0.11 (0.03)	0.10 (0.04)	0.06 (0.03)	0.08 (0.03)	0.01 (0.03)	0.04 (0.03)	0.03 (0.03)	0.01 (0.03)	0.06 (0.03)
ICW: Podcast take-up	0.15 (0.03)	0.02 (0.02)					0.03 (0.02)		0.02 (0.03)	0.03 (0.02)	0.06 (0.02)		0.06 (0.03)
Behavior: Stayed home	0.01 (0.02)				0.05 (0.02)		0.02 (0.02)		0.02 (0.02)		0.02 (0.01)	0.02 (0.02)	
Behavior: Wore mask	0.07 (0.02)	0.05 (0.02)		0.01 (0.02)	0.01 (0.01)						0.01 (0.01)		
Behavior: Visited indoors (reversed)	0.02 (0.02)	0.02 (0.02)	0.01 (0.02)	0.01 (0.02)	0.02 (0.02)	0.02 (0.02)						0.04 (0.02)	0.01 (0.02)
Gender: Female	0.05 (0.04)	0.03 (0.04)	0.01 (0.03)	0.03 (0.04)							0.01 (0.03)	0.11 (0.03)	0.01 (0.04)
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)	0.03 (0.01)			0.02 (0.01)		0.02 (0.01)
Education: University	0.03 (0.02)	0.07 (0.02)	0.07 (0.02)	0.04 (0.02)	0.01 (0.02)	0.04 (0.02)	0.02 (0.02)		0.02 (0.02)	0.02 (0.02)		0.02 (0.02)	0.01 (0.02)
Province: Free State	0.04 (0.02)	0.04 (0.02)		0.02 (0.02)			0.03 (0.01)	0.02 (0.01)		0.02 (0.01)	0.02 (0.01)	0.01 (0.02)	
Province: Gauteng	0.06 (0.03)	0.01 (0.02)		0.02 (0.02)		0.02 (0.02)	0.00 (0.02)	0.00 (0.02)	0.02 (0.02)			0.00 (0.03)	0.04 (0.02)
Province: KwaZulu-Natal	0.04 (0.02)	0.04 (0.02)		0.03 (0.02)	0.00 (0.02)						0.01 (0.02)	0.02 (0.02)	0.02 (0.02)
Province: Limpopo	0.05 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.00 (0.02)						0.02 (0.02)	0.01 (0.02)	
Province: Mpumalanga	0.03 (0.02)		0.01 (0.02)		0.04 (0.02)								0.03 (0.02)
Province: North West	0.01 (0.02)	0.03 (0.02)	0.02 (0.02)	0.03 (0.02)			0.04 (0.02)				0.01 (0.02)	0.02 (0.02)	
Province: Western Cape	0.02 (0.02)		0.02 (0.02)		0.03 (0.02)			0.03 (0.02)	0.03 (0.02)	0.03 (0.02)	0.05 (0.01)	0.03 (0.02)	0.00 (0.02)
Locality: Peri-urban	0.02 (0.03)			0.06 (0.02)				0.01 (0.02)	0.01 (0.03)				
Locality: Rural	0.07 (0.03)	0.03 (0.02)	0.01 (0.02)			0.01 (0.02)			0.02 (0.03)				0.07 (0.02)
ICW: Verify challenge	0.05 (0.02)		0.02 (0.01)		0.01 (0.02)		0.01 (0.02)		0.02 (0.02)			0.01 (0.02)	0.02 (0.01)
ICW: Consume news from close friends	0.02 (0.02)	0.08 (0.02)	0.05 (0.02)	0.03 (0.02)	0.02 (0.02)	0.02 (0.02)	0.07 (0.02)		0.06 (0.02)	0.03 (0.02)	0.01 (0.02)		0.05 (0.02)
ICW: Consume social media	0.03 (0.03)	0.06 (0.02)	0.04 (0.02)	0.09 (0.02)	0.09 (0.02)		0.05 (0.02)		0.14 (0.02)		0.02 (0.02)	0.02 (0.02)	0.00 (0.02)
ICW: Consume traditional media	0.04 (0.02)	0.03 (0.02)	0.04 (0.02)	0.03 (0.02)	0.06 (0.01)	0.02 (0.02)	0.03 (0.02)			0.02 (0.02)	0.01 (0.01)	0.04 (0.02)	0.05 (0.02)
ICW: Consume WhatsApp	0.05 (0.02)	0.05 (0.02)	0.07 (0.02)	0.03 (0.02)	0.03 (0.02)	0.06 (0.02)	0.02 (0.02)	0.02 (0.02)		0.03 (0.02)	0.13 (0.02)	0.03 (0.02)	0.10 (0.02)
ICW: COVID-19 beliefs and behavior	0.07 (0.03)	0.01 (0.02)	0.03 (0.02)			0.02 (0.02)		0.03 (0.01)	0.05 (0.02)	0.05 (0.02)		0.16 (0.02)	0.01 (0.02)
ICW: Non-WCW podcast take-up	0.05 (0.02)		0.01 (0.01)	0.07 (0.02)	0.04 (0.02)	0.03 (0.02)			0.03 (0.02)		0.07 (0.02)		0.07 (0.02)
ICW: Misinformation harmful	0.05 (0.02)	0.01 (0.02)	0.03 (0.01)	0.01 (0.02)					0.02 (0.02)		0.01 (0.01)	0.01 (0.02)	0.07 (0.02)
ICW: Sharing	0.01 (0.02)	0.00 (0.02)	0.04 (0.02)	0.04 (0.02)	0.07 (0.02)	0.04 (0.02)		0.08 (0.02)	0.02 (0.02)	0.02 (0.02)	0.16 (0.02)	0.03 (0.02)	0.03 (0.02)
ICW: Trust organizations	0.03 (0.02)	0.02 (0.02)	0.06 (0.02)	0.01 (0.02)			0.01 (0.02)				0.01 (0.02)	0.01 (0.02)	0.04 (0.02)
ICW: Trust traditional media	0.02 (0.02)	0.04 (0.02)		0.04 (0.02)	0.06 (0.02)	0.01 (0.02)	0.04 (0.02)			0.02 (0.02)	0.01 (0.02)	0.04 (0.02)	0.07 (0.02)
ICW: Trust WhatsApp	0.01 (0.02)			0.05 (0.02)	0.02 (0.02)		0.01 (0.02)	0.03 (0.02)		0.01 (0.02)	0.05 (0.02)		0.02 (0.02)
Get news from: Other social media	0.01 (0.02)			0.04 (0.02)	0.06 (0.02)			0.06 (0.01)	0.04 (0.02)				0.03 (0.02)
How often listens to podcasts	0.08 (0.03)		0.03 (0.02)	0.01 (0.02)				0.02 (0.02)		0.04 (0.02)		0.02 (0.02)	
Age: 18-24	0.06 (0.02)		0.09 (0.02)	0.03 (0.02)	0.21 (0.02)	0.02 (0.02)	0.01 (0.02)		0.05 (0.02)	0.05 (0.02)	0.02 (0.02)	0.14 (0.02)	0.18 (0.02)
Age: 25-34	0.09 (0.02)	0.01 (0.02)		0.03 (0.02)	0.10 (0.02)							0.07 (0.02)	0.12 (0.02)
To verify: Ask in person (reversed)	0.04 (0.02)	0.02 (0.02)		0.01 (0.02)					0.02 (0.02)		0.03 (0.02)	0.01 (0.02)	
To verify: Use FC	0.01 (0.02)	0.05 (0.02)		0.03 (0.02)	0.04 (0.02)	0.02 (0.02)	0.08 (0.02)	0.04 (0.02)		0.03 (0.02)	0.02 (0.02)	0.02 (0.02)	
How often verify	0.02 (0.02)	0.03 (0.02)	0.05 (0.01)	0.03 (0.02)	0.03 (0.02)	0.06 (0.02)	0.08 (0.02)				0.03 (0.01)	0.01 (0.02)	0.01 (0.02)
To verify: Ask others on WA (reversed)	0.01 (0.02)		0.01 (0.01)		0.05 (0.02)				0.02 (0.01)		0.03 (0.02)		0.05 (0.02)
To verify: Post on social media (reversed)	0.04 (0.02)		0.03 (0.01)	0.01 (0.02)				0.07 (0.02)	0.03 (0.02)				
Locality: Urban		0.05 (0.04)	0.02 (0.04)	0.01 (0.04)	0.00 (0.04)			0.06 (0.04)			0.03 (0.04)	0.03 (0.04)	
ICW: Trust close friends		0.02 (0.02)	0.01 (0.01)	0.04 (0.02)	0.03 (0.02)		0.01 (0.02)		0.06 (0.02)			0.03 (0.02)	0.03 (0.02)
ICW: Verification knowledge (part 2)		0.02 (0.03)		0.01 (0.03)	0.01 (0.03)			0.01 (0.02)		0.02 (0.02)	0.05 (0.03)	0.04 (0.03)	0.01 (0.02)
Listens to WCW		0.03 (0.02)	0.03 (0.02)	0.00 (0.01)	0.03 (0.02)		0.02 (0.02)		0.02 (0.02)		0.01 (0.02)	0.05 (0.01)	0.02 (0.02)
Age: 35-44		0.02 (0.02)	0.03 (0.01)								0.05 (0.02)		
To verify: Ask family on WA (reversed)		0.01 (0.02)	0.04 (0.01)				0.04 (0.02)		0.03 (0.02)			0.03 (0.02)	
To verify: Use internet		0.08 (0.02)	0.02 (0.02)	0.06 (0.02)	0.07 (0.02)	0.03 (0.02)	0.09 (0.02)		0.02 (0.02)	0.04 (0.02)	0.01 (0.02)	0.01 (0.02)	0.04 (0.02)
Province: Eastern Cape			0.02 (0.01)	0.02 (0.02)	0.02 (0.01)		0.02 (0.02)	0.04 (0.01)	0.01 (0.02)	0.03 (0.02)	0.02 (0.01)	0.04 (0.02)	
ICW: Trust social media					0.06 (0.02)			0.20 (0.02)	0.06 (0.02)		0.06 (0.02)	0.01 (0.02)	0.07 (0.02)
ICW: Verification knowledge						0.11 (0.02)							
ICW: Active verification										0.18 (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	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	1.00
R^2	0.23	0.26	0.14	0.13	0.16	0.10	0.13	0.17	0.14	0.14	0.23	0.14	0.20
Observations	4543	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 covariate coefficients)

Figure	4a	4b	4c	5a	5	6a	6b	6c	7a	7b	7c	8a	8b
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Placebo incentives	0.33 (0.05)	0.13 (0.05)	0.07 (0.05)	0.05 (0.05)	0.01 (0.05)	0.05 (0.05)	0.04 (0.05)	0.03 (0.05)	0.02 (0.05)	0.03 (0.05)	0.01 (0.04)	0.03 (0.05)	0.10 (0.05)
Text information	0.01 (0.06)	0.34 (0.06)	0.24 (0.06)	0.11 (0.06)	0.11 (0.06)	0.18 (0.06)	0.02 (0.06)	0.14 (0.06)	0.07 (0.06)	0.13 (0.06)	0.09 (0.05)	0.16 (0.06)	0.09 (0.06)
Short podcast	0.64 (0.05)	0.39 (0.05)	0.24 (0.04)	0.03 (0.04)	0.05 (0.04)	0.12 (0.05)	0.07 (0.04)	0.02 (0.04)	0.02 (0.04)	0.04 (0.04)	0.02 (0.04)	0.03 (0.04)	0.11 (0.04)
Long podcast	0.65 (0.05)	0.38 (0.05)	0.17 (0.04)	0.00 (0.05)	0.13 (0.04)	0.03 (0.05)	0.00 (0.05)	0.07 (0.04)	0.01 (0.04)	0.02 (0.04)	0.01 (0.04)	0.01 (0.05)	0.04 (0.05)
Empathetic podcast	0.66 (0.05)	0.51 (0.05)	0.20 (0.04)	0.14 (0.04)	0.16 (0.04)	0.11 (0.05)	0.12 (0.05)	0.14 (0.04)	0.04 (0.04)	0.04 (0.05)	0.07 (0.04)	0.05 (0.04)	0.02 (0.05)
ICW: Podcast take-up	0.15 (0.03)	0.01 (0.03)			0.00 (0.03)		0.03 (0.02)		0.02 (0.03)	0.03 (0.02)	0.05 (0.02)		0.06 (0.03)
Behavior: Stayed home	0.01 (0.02)	0.00 (0.02)			0.05 (0.02)		0.02 (0.02)		0.02 (0.02)		0.04 (0.02)	0.02 (0.02)	
Behavior: Wore mask	0.07 (0.02)	0.05 (0.02)		0.01 (0.02)	0.01 (0.01)						0.03 (0.02)		0.02 (0.02)
Behavior: Visited indoors (reversed)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.01 (0.02)	0.02 (0.02)	0.02 (0.02)						0.04 (0.02)	0.01 (0.02)
Gender: Female	0.06 (0.04)	0.03 (0.04)	0.01 (0.03)	0.03 (0.04)							0.01 (0.04)	0.11 (0.03)	0.01 (0.04)
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)	0.01 (0.01)		0.02 (0.01)		0.02 (0.01)
Education: University	0.03 (0.02)	0.06 (0.02)	0.07 (0.02)	0.04 (0.02)	0.01 (0.02)	0.04 (0.02)	0.02 (0.02)		0.02 (0.02)	0.02 (0.02)		0.01 (0.02)	0.01 (0.02)
Province: Eastern Cape	0.04 (0.03)	0.00 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (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)			0.03 (0.01)	0.02 (0.01)		0.02 (0.01)	0.02 (0.01)		
Province: Gauteng	0.12 (0.06)	0.01 (0.02)	0.00 (0.02)	0.02 (0.02)		0.02 (0.02)	0.01 (0.02)	0.00 (0.02)	0.02 (0.02)			0.02 (0.02)	0.04 (0.02)
Province: KwaZulu-Natal	0.09 (0.05)	0.04 (0.02)		0.03 (0.02)	0.00 (0.02)						0.01 (0.02)		0.02 (0.02)
Province: Limpopo	0.08 (0.04)	0.03 (0.02)	0.02 (0.02)	0.02 (0.02)	0.00 (0.02)						0.02 (0.01)	0.03 (0.02)	
Province: Mpumalanga	0.06 (0.04)	0.00 (0.02)	0.01 (0.02)		0.04 (0.02)					0.01 (0.02)			0.03 (0.02)
Province: North West	0.04 (0.04)	0.04 (0.02)	0.02 (0.02)	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 (0.02)			0.03 (0.02)	0.03 (0.02)	0.03 (0.02)	0.05 (0.01)	0.01 (0.02)	0.00 (0.02)
Locality: Peri-urban	0.02 (0.03)	0.04 (0.02)		0.14 (0.05)				0.01 (0.02)	0.01 (0.03)				
Locality: Rural	0.07 (0.03)		0.01 (0.02)	0.09 (0.05)		0.01 (0.02)			0.02 (0.03)				0.07 (0.02)
ICW: Verify challenge	0.05 (0.02)	0.00 (0.02)	0.02 (0.01)		0.01 (0.02)		0.01 (0.02)		0.02 (0.02)				0.02 (0.01)
ICW: Consume news from close friends	0.02 (0.02)	0.08 (0.02)	0.05 (0.02)	0.03 (0.02)	0.02 (0.02)	0.02 (0.02)	0.07 (0.02)		0.06 (0.02)	0.04 (0.02)	0.01 (0.02)		0.05 (0.02)
ICW: Consume social media	0.03 (0.03)	0.08 (0.02)	0.04 (0.02)	0.09 (0.02)	0.09 (0.02)		0.05 (0.02)		0.14 (0.02)		0.05 (0.02)	0.02 (0.02)	0.01 (0.02)
ICW: Consume traditional media	0.05 (0.02)	0.03 (0.02)	0.04 (0.02)	0.03 (0.02)	0.06 (0.01)	0.02 (0.02)	0.03 (0.02)			0.02 (0.02)	0.01 (0.01)	0.04 (0.02)	0.05 (0.02)
ICW: Consume WhatsApp	0.05 (0.02)	0.04 (0.02)	0.07 (0.02)	0.03 (0.02)	0.03 (0.02)	0.06 (0.02)	0.02 (0.02)	0.02 (0.02)	0.00 (0.02)	0.03 (0.02)	0.13 (0.02)	0.03 (0.02)	0.10 (0.02)
ICW: COVID-19 beliefs and behavior	0.08 (0.03)		0.03 (0.02)			0.02 (0.02)		0.03 (0.01)	0.05 (0.02)	0.05 (0.02)	0.04 (0.02)	0.16 (0.02)	0.02 (0.02)
ICW: Non-WCW podcast take-up	0.05 (0.02)	0.01 (0.02)	0.02 (0.01)	0.07 (0.02)	0.04 (0.02)	0.03 (0.02)			0.03 (0.02)		0.07 (0.02)		0.07 (0.02)
ICW: Misinformation harmful	0.05 (0.02)	0.01 (0.02)	0.03 (0.01)	0.01 (0.02)					0.02 (0.02)		0.01 (0.01)		0.07 (0.02)
ICW: Sharing	0.01 (0.02)	0.00 (0.02)	0.04 (0.02)	0.04 (0.02)	0.07 (0.02)	0.03 (0.02)		0.08 (0.02)	0.02 (0.02)	0.02 (0.02)	0.16 (0.02)	0.03 (0.02)	0.03 (0.02)
ICW: Trust close friends	0.01 (0.02)	0.03 (0.02)	0.01 (0.01)	0.04 (0.02)	0.03 (0.02)		0.01 (0.02)		0.06 (0.02)			0.02 (0.02)	0.03 (0.02)
ICW: Trust organizations	0.03 (0.02)	0.03 (0.02)	0.06 (0.02)	0.01 (0.02)			0.01 (0.02)				0.01 (0.02)	0.01 (0.02)	0.04 (0.02)
ICW: Trust traditional media	0.02 (0.02)	0.04 (0.02)		0.04 (0.02)	0.06 (0.02)	0.01 (0.02)	0.04 (0.02)			0.02 (0.02)	0.01 (0.02)	0.04 (0.02)	0.07 (0.02)
ICW: Trust WhatsApp	0.02 (0.02)	0.02 (0.02)		0.05 (0.02)	0.02 (0.02)		0.01 (0.02)	0.03 (0.02)		0.01 (0.02)	0.05 (0.02)		0.02 (0.02)
ICW: Verification knowledge (part 2)	0.02 (0.03)	0.01 (0.05)		0.01 (0.03)	0.03 (0.03)			0.01 (0.02)		0.00 (0.02)	0.05 (0.03)	0.05 (0.03)	0.01 (0.02)
Get news from: Other social media	0.02 (0.02)	0.03 (0.02)		0.04 (0.02)	0.06 (0.02)			0.07 (0.01)	0.04 (0.02)		0.04 (0.02)		0.03 (0.02)
How often listens to podcasts	0.07 (0.03)		0.03 (0.02)					0.02 (0.02)		0.04 (0.02)		0.02 (0.02)	
Age: 18-24	0.07 (0.02)	0.01 (0.04)	0.09 (0.02)	0.03 (0.02)	0.20 (0.02)	0.02 (0.02)	0.01 (0.02)		0.05 (0.02)	0.05 (0.02)	0.02 (0.02)	0.17 (0.04)	0.18 (0.02)
Age: 25-34	0.09 (0.02)	0.01 (0.04)		0.03 (0.02)	0.10 (0.02)							0.10 (0.04)	0.12 (0.02)
To verify: Ask in person (reversed)	0.03 (0.02)	0.01 (0.02)		0.01 (0.02)					0.02 (0.02)	0.03 (0.02)	0.03 (0.02)		
To verify: Use FC	0.00 (0.02)	0.05 (0.03)		0.03 (0.02)	0.03 (0.02)	0.02 (0.02)	0.08 (0.02)	0.06 (0.02)		0.04 (0.02)	0.02 (0.02)	0.02 (0.02)	
How often verify	0.01 (0.02)	0.03 (0.02)	0.05 (0.01)	0.03 (0.02)	0.03 (0.02)	0.06 (0.02)	0.07 (0.02)				0.04 (0.01)	0.01 (0.02)	0.01 (0.02)
To verify: Ask others on WA (reversed)	0.00 (0.02)	0.00 (0.02)	0.01 (0.01)		0.04 (0.02)				0.02 (0.01)		0.03 (0.02)		0.05 (0.02)
To verify: Post on social media (reversed)	0.04 (0.02)		0.03 (0.01)	0.01 (0.02)				0.08 (0.02)	0.03 (0.02)				
Locality: Urban		0.01 (0.04)	0.02 (0.04)	0.10 (0.07)	0.00 (0.04)			0.06 (0.04)			0.02 (0.04)	0.03 (0.04)	
ICW: Trust social media		0.03 (0.02)			0.06 (0.02)			0.20 (0.02)	0.06 (0.02)		0.06 (0.02)		0.07 (0.02)
Listens to WCW		0.03 (0.02)	0.03 (0.02)		0.02 (0.02)		0.02 (0.02)		0.02 (0.02)		0.02 (0.02)	0.05 (0.01)	0.02 (0.02)
Age: 35-44		0.02 (0.03)	0.03 (0.01)								0.05 (0.02)	0.03 (0.03)	
To verify: Ask family on WA (reversed)		0.01 (0.02)	0.04 (0.01)		0.02 (0.02)		0.04 (0.02)		0.03 (0.02)			0.03 (0.02)	
To verify: Use internet		0.08 (0.03)	0.02 (0.02)	0.06 (0.02)	0.06 (0.02)	0.03 (0.02)	0.09 (0.02)	0.03 (0.02)	0.02 (0.02)	0.04 (0.02)	0.01 (0.02)	0.00 (0.02)	0.05 (0.02)
ICW: Verification knowledge						0.11 (0.02)							
ICW: Active verification										0.18 (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	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	1.00
R^2	0.26	0.26	0.14	0.13	0.16	0.10	0.13	0.17	0.14	0.14	0.23	0.14	0.20
Observations	4543	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 F16: Aggregating ICW indexes from each figure and correcting for multiple comparisons

	Index of Figure 4 outcomes		Index of Figure 5 outcomes		Index of Figure 6 outcomes		Index of Figure 7 outcomes		Index of Figure 8 outcomes	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A. Pooled estimation										
Placebo incentives	0.264	0.281	0.014	0.028	0.060	0.074	-0.029	-0.012	0.024	0.041
	(0.050)	(0.048)	(0.051)	(0.049)	(0.051)	(0.049)	(0.050)	(0.048)	(0.049)	(0.047)
	[0.000]	[0.000]	[0.390]	[0.285]	[0.121]	[0.068]	[0.552]	[0.795]	[0.307]	[0.191]
Pooled treatment	0.639	0.646	0.103	0.110	0.142	0.139	0.004	0.000	0.034	0.047
	(0.035)	(0.034)	(0.035)	(0.034)	(0.036)	(0.034)	(0.035)	(0.034)	(0.034)	(0.033)
	[0.000]	[0.000]	[0.003]	[0.000]	[0.000]	[0.000]	[0.456]	[0.996]	[0.204]	[0.100]
B. Disaggregated estimation										
Placebo incentives	0.265	0.282	0.015	0.028	0.060	0.074	0.029	0.012	0.024	0.042
	(0.050)	(0.048)	(0.051)	(0.049)	(0.051)	(0.049)	(0.050)	(0.048)	(0.049)	(0.047)
	[0.000]	[0.000]	[0.387]	[0.287]	[0.119]	[0.068]	[0.553]	[0.794]	[0.309]	[0.189]
Text information	0.283	0.310	0.144	0.141	0.196	0.189	0.019	0.018	0.139	0.163
	(0.059)	(0.055)	(0.061)	(0.059)	(0.064)	(0.061)	(0.061)	(0.060)	(0.058)	(0.056)
	[0.000]	[0.000]	[0.009]	[0.008]	[0.001]	[0.000]	[0.376]	[0.385]	[0.009]	[0.002]
Short podcast	0.686	0.691	0.042	0.045	0.134	0.123	0.043	0.049	0.085	0.090
	(0.048)	(0.047)	(0.046)	(0.045)	(0.047)	(0.046)	(0.045)	(0.044)	(0.044)	(0.043)
	[0.000]	[0.000]	[0.183]	[0.160]	[0.002]	[0.004]	[0.344]	[0.259]	[0.027]	[0.018]
Long podcast	0.649	0.654	0.058	0.080	0.051	0.060	0.001	0.008	0.001	0.014
	(0.050)	(0.049)	(0.047)	(0.045)	(0.048)	(0.046)	(0.046)	(0.044)	(0.046)	(0.045)
	[0.000]	[0.000]	[0.106]	[0.038]	[0.142]	[0.097]	[0.491]	[0.427]	[0.991]	[0.378]
Empathetic podcast	0.742	0.744	0.194	0.195	0.217	0.212	0.048	0.034	0.033	0.018
	(0.049)	(0.048)	(0.046)	(0.044)	(0.048)	(0.046)	(0.046)	(0.045)	(0.046)	(0.044)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.150]	[0.220]	[0.464]	[0.677]
Controls	×	✓	×	✓	×	✓	×	✓	×	✓
Directional hypothesis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control Mean	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
R^2	0.21	0.28	0.09	0.16	0.08	0.16	0.08	0.15	0.08	0.15
Observations	4543	4543	4543	4543	4543	4543	4543	4543	4543	4543

Notes: Outcomes are standardized indexes of all ICW indexes in a given figure (see column headers). ICW indexes are reversed to ensure hypotheses share the same expected direction. 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. p-values in square brackets adjust for multiple testing using Benjamini-Hochberg correction both across treatment coefficients and across outcomes (holding fixed whether the specification includes controls or not).

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:

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

where Y_{ib} is an outcome for respondent i from block b 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.