The Effects of Sustained Exposure to Fact-checking Information: Evidence from a Field Experiment on Twitter

Tiago Ventura [*a, c], Kevin Aslett[b,c], Felicia Loecherbach[c, d], Joshua A. Tucker[c, e], and Solomon Messing[c]

[a]McCourt School of Public Policy, Georgetown University
[b]School of Politics, Security, and International Affairs, University of Central Florida
[c]Center for Social Media and Politics, New York University
[d]Amsterdam School of Communication Research, University of Amsterdam
[e]Department of Politics, New York University, New York

Abstract

Social media companies and civic society rely heavily on fact-checking to counter misinformation. While numerous studies have shown the efficacy of single-shot corrective interventions, the effects of sustained exposure to fact-checking information in a realistic social media environment have yet to be tested. In this study, we conduct a one-month field intervention implemented on a widely used social media platform to analyze the causal effect of substantially increasing users' exposure to fact-checking accounts and content on resilience to misinformation and attitudinal outcomes. In our design, Twitter users will be randomly assigned to an intervention group that will have a new timeline in their accounts composed of a pre-curated list of fact-checking organizations added to the top of their Twitter feeds, and a control group where nothing is added. Over a four-week period, participants' compliance with the intervention will be consistently assessed, and two survey waves will measure outcomes of interest.

[*]Corresponding author. Email: tv186@georgetown.edu

Introduction

The spread of online misinformation on social media and its potential effects on citizens' attitudes and behaviors have recently become a subject of widespread concern among academics, policy-makers, and the media. Around the world, scholarly and journalistic accounts have shown that coordinated misinformation campaigns on social media have attempted to influence elections, turnout and voting choices (1, 2, 3), increased vaccine hesitancy during the Covid-19 pandemic (4), generated distrust on scientific information about climate change (5) and even rendered episodes of offline violence against minority groups (6, 7). In this context, social media companies, researchers, and policymakers have developed strategies to pre- and debunk false claims spreading on social media platforms in collaboration with fact-checkers: Professional, independent organizations that check suspect claims spreading online and report on their truthfulness. This investment has been supported by extensive literature that has attested to fact-checking corrections' substantive positive effect on subjects' capacity to improve their ability to identify true and false information (8, 9, 10, 11, 12).

This emerging literature varies in context, type, and timing of interventions. Yet their evidence has, by and large, been collected through one-shot survey experiments. For example, we conducted a simple meta-analysis using a database of randomized controlled interventions focusing on countering misinformation beliefs recently published by USAID (13) and found that only 13% (twenty-one studies) relied on field designs, while all the rest used survey design-based experiments. While survey experiments benefit from strong internal validity, allowing for precise measures about the cognitive mechanism behind beliefs for misinformation, they lack ecological validity, particularly the capacity to understand the effects of corrective intervention in naturalistic social media environments. For example, two key issues might emerge within these designs: (1) Demand effects: Respondents might behave differently in a one-shot survey

experiment than outside of this controlled research setting ; (2) Selection effects: Outside of a controlled research setting, respondents might select out of fact-checking corrections to avoid the cognitive costs of being exposed to counter-attitudinal corrections. The value of large-scale field experiments to understand the causal effects of social consumption has recently been illustrated through research-industry collaborations such as the U.S. 2020 Facebook and Instagram Election Studies, demonstrating the nuances in effects of, for example changing social media timeline rankings (14, 15) on media diet and attitudinal changes. Similarly, past research on fact-checking on Twitter and especially the crowd fact-checking annotation program Birdwatch has occured internally or in cooperation with research teams (16, 17).

In this Registered Report, we propose an online experiment that will measure the effect of sustained exposure to fact-checks in the field using Twitter's inherent features that can be easily implemented without any industry cooperation [1]. Even though we focus on Twitter, our intervention can be easily extrapolated and implemented to other micro-blogging platforms, such as BlueSky and Meta-owned Threads, for example. We design an intervention that creates an additional timeline on respondents' Twitter feeds formed by eight reputable fact-checking organizations with active Twitter accounts. Our intervention focuses on substantially increasing users' exposure to fact-checking accounts on a widely used social media platform, Twitter. This new timeline is added at the side of the home timeline on Twitter and allows users to seamlessly swipe left and right across their home timeline and the recently added timeline. To avoid demand effects, we label the intervention as a **Media Timeline** and explain to participants that the timeline comprises media organizations that verify the accuracy of online news. The intervention is implemented with very little disturbance in the respondents' online environment. Through an intervention using a new timeline with many fact-checking accounts rather than a single news source in particular, our design preserves respondents' agency over (selective

[1] Twitter has recently changed the legal company name to X. Given the platform is still largely known by users and general population as Twitter, we prefer to use Twitter throughout the manuscript

exposure to) the information shared in the timeline. This design feature helps address concerns that forced-exposure designs can exaggerate effect sizes (18) by preserving participants' ability to selectively expose themselves to the different organizations added to the timeline. Figure 1 visually presents the proposed intervention.

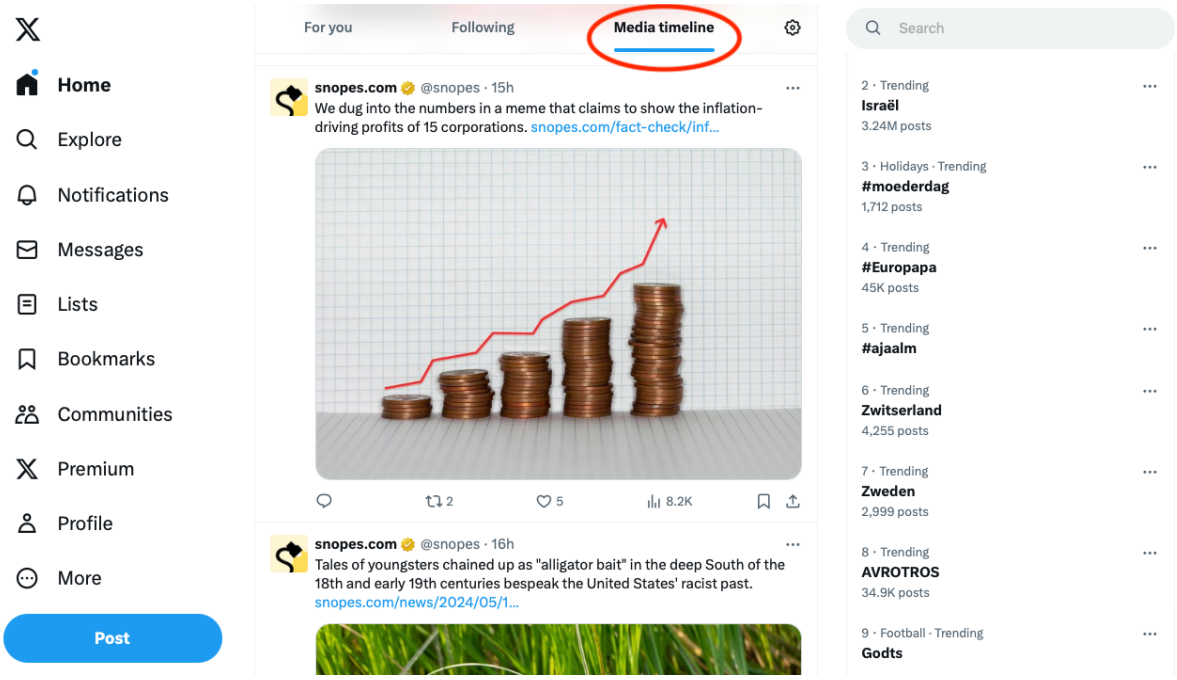


Figure 1 Example of the Treatment Intervention

To implement the intervention, we developed an App asking for users' consent to post requests to their accounts using the Twitter API. Participants in the control group will be asked to continue to use Twitter as normal and will not see any changes on their Twitter feed. The treatment period will last for four weeks, and outcomes will be measured using survey data collected at the end of the study (See Appendix B for a visual demonstration of the new timeline). Because our intervention sought to test whether sustained exposure to fact-checks affects misinformation resilience and attitudes, we provide incentives to increase participants' exposure to the content posted in the timeline. In particular, we incentivize participants to scroll down

through their new timeline and send us daily screenshots of tweets posted by the fact-checking organizations added to the timeline. To ensure that our results are not driven by an increase in time spent on the platform, we will also offer incentives to the control group to send us daily screenshots of a tweet they found most informative that appeared on their Twitter feed on any given day. Our incentive structure is commensurate with other online field experiments, for example, measuring the effects of exposure counter-attitudinal bots on Twitter (19), increasing sustained consumption of fact-checks on WhatsApp (20), and switching media habits among TV viewers (21) that provide similar incentives to increase treatment reception among participants.

Ahead of past critical elections, Twitter deployed similar strategies to increase the uptake of credible information on the platform. For example, for the recent Presidential election in Brazil 2022, Twitter created a list of Fact-Checking organizations working on issues related to the elections and incentivized through communication campaigns and advertisements in the platform users to follow and pin the list on their timelines. A similar tool was provided to users during the mid-term elections in the United States in 2022.

Our design uses a similar intervention as those deployed previously by Twitter, combined with a strong incentive structure for participants to engage with tweets from fact-checking organizations. While we do not propose here to perform a policy evaluation of these previous interventions, our design builds upon this easy-to-implement tool available for all Twitter users to measure the effects of sustained exposure to fact-checking in a realistic and ecologically valid manner. Moreover, as the vast majority of the literature attests to a substantive effect of debunking interventions on accuracy beliefs (8, 9, 10, 11, 12), a pivotal goal of our design is to assess the robustness of these findings to when exposure to fact-checking occurs in the field, on realistic social media environments. To do so, our design relies on a high-dosage intervention, incentivizing participants to continuously engage with fact-checking organizations, purpose-

fully to avoid the risk of null effects due to a weak instrument. In this way, our intervention is capable of observing precise effects, including a null result, which will provide important insights about the external validity of previous findings based mostly on one-shot survey interventions. To guarantee our findings contribute to the literature, we propose a high dosage experiment capable of observing effects, including a null, that is substantively meaningful because it would suggest that the actual effect, without incentives, for example, is so small that it cannot be material in the real world/policy world.

Our design continuously increases individuals' consumption of fact-checking corrections through two distinct paths. On one side, participants might engage with posts assessing the accuracy of false (and true) information circulating online and offline. In addition, through a sustained and repeated process of engaging with fact-check accounts, participants also engage with posts forewarning individuals about potential misinformation before they encounter it and inoculating them against these falsehoods. These paths are known in the scholarly work as debunking and prebunking interventions.

The psychological principles behind most misinformation interventions align with dual process models for how attitudes and persuasion operate. In these models, human cognition operates either through fast/automatic processing, or deliberative, analytic processes. Through distinct psychological mechanisms, either prebunking and debunking interventions intend to incentivize individuals to rely on the latter model and, by doing so, increase the importance of accuracy motivations on belief formation (22, 23) in contrast to fast, automatic processing, which is often associated to directional-motivated reasoning (24, 25) For example, debunking false rumors and other forms of corrective information tends to increase the accuracy motivations of people's beliefs (26) and online behavior (23), particularly in the contexts of repeated exposure, which might increase familiarity with corrections, create greater fluency with true claims, and reduce novelty effects of misinformation rumors (27, 28) Moreover, recent stud-

ies show how inoculating users with fact-checks confers psychological resistance and enhances critical thinking skills, making individuals less susceptible to false claims (29). Moreover, these preemptive interventions can also improve consumers' media literacy by providing common fingerprints about misinformation rumors and improving users' skills to spot them on future episodes of misinformation exposure (30).

Critically, our experiment focuses on two broader sets of outcomes. First, we analyze the effects of sustained exposure to fact-checking on users' resilience to misinformation. Second, we focus on the attitudinal effects of fact-checking exposure on trust in mainstream media and political and social media cynicism. Specifically, our first research question asks: To what extent does sustained exposure to fact-checking information in a real-world environment on social media affect misinformation resilience? For this study, the term misinformation resilience subsumes three different concepts: 1) Ability to discern between true and false information (factual discernment) 2) Understanding and relying on proper methods to verify the accuracy of news (misinformation literacy) and 3) Tendency to engage with conspiracist explanations (conspiratorial predispositions).

Previous research has found positive effects for prebunking (29, 31) debunking interventions (8, 9, 10, 11, 12, 11), as well as other similar approaches like literacy interventions (32), and news quality labels on factual discernment (33, 32). However, in this literature, while effects are generally positive on reducing misperceptions, they are often small in magnitude and attenuate quickly over time (26). Our first hypothesis assesses the robustness of these findings to a field design, using an ecologically valid intervention, leveraging both prebunking and debunking strategies, and with incentives for participants to continuously engage with fact-check content online. Furthermore, sustained exposure to professional fact-checks could improve the reader's misinformation literacy by providing them with proper strategies to investigate future suspect claims. These effects tend to be more prominent when leveraging fact-checking as a pre-

bunking mechanism, which can help participants to build misinformation resilience (29). Our second hypothesis tests for these effects. Lastly, exposure to professional fact-checks may also lower conspiratorial predispositions, because many fact-checks are aimed at debunking larger conspiracy theories such as those around unfounded allegations of election fraud or Covid-19 related conspiracies (34). There has been no research, to our knowledge, that has tested whether fact-checking can be effective for countering conspiratorial predispositions, but it is one of the few (low-cost) strategies available (35, 36).

Our three pre-registered hypotheses that investigate the effect of real-world sustained exposure to fact-checking on social media news feeds on three different levels of misinformation resilience are listed below:

- Hypothesis 1: Exposure to fact-checking content will increase subjects' ability to accurately discern between TRUE and FALSE information compared to users that are using Twitter as usual

- Hypothesis 2: Exposure to fact-checking content will increase subjects' misinformation literacy compared to users who are using Twitter as usual.

- Hypothesis 3: Exposure to fact-checking content will decrease subjects' conspiratorial predispositions compared to users who are using Twitter as usual.

Given that our study plans to substantially increase exposure to fact-checks, we also seek to test whether this increase affects attitudes towards the subjects of these fact-checks: mainstream media, politicians and political organizations, and social media posts. To this end, we focus on three different measurements: trust in the mainstream media, cynicism towards politics, and cynicism towards information circulating on social media. We use cynicism as concept as it describes a more abstract, generalized negative feeling towards politics and information

circulating on social media as opposed to specific mistrust towards particular agents (politicians, political groups, social media influences, online news pages) (37). We ask: To what extent does sustained exposure to fact-checks on social media affect attitudes towards media organizations, politicians, and social media content?

Frequent correction of news items or political information could reduce the credibility of the producer of that information, such as the media or politicians or the platform that amplifies that information. Previous work has suggested that high levels of exposure to fact-checking could make individuals more cynical and likely to distrust all media (38) and recent studies provide evidence of this. For example, (39) found that exposure to political fact-checks decreased trust in news.[2] Similarly, exposure to the discourse around fake news reduces trust in news media in general (40). But, the media is not the only group often fact-checked. Fact-checking organizations also post fact-checks about politicians, partisan media, and political organizations, and frequent fact-checks of statements made by political leaders may also, like the media, lead to lower levels of trust and greater cynicism towards politicians and political institutions in general (41, 42). Moreover, often viral social media posts, without a clear source attribution, are also fact-checked by professional fact-checking organizations, so this suggests that exposure to these professional fact-checks should also increase cynicism of content circulating on social media through the same mechanism as described previously. Frequent corrections of information posted on social media platforms will likely further reinforce high levels of skepticism individuals already hold for information they see on social media (43), and as shown in (44) exposure to debunking messages reduce beliefs for both true and false news, suggesting that sustained exposure to fact-checking efforts can also increase distrust of legitimate news and information. Based on these arguments, we expect:

- Hypothesis 4: Exposure to fact-checking content will decrease subjects' trust in the main-

[2]Exposure corrections reduced trust by 6-8% compared to those who saw uncorrected misinformation.

stream media compared to users that are using Twitter as usual.

- Hypothesis 5: Exposure to fact-checking content will increase subjects' cynicism towards politics compared to users that are using Twitter as usual.

- Hypothesis 6: Exposure to fact-checking content will increase subjects' cynicism towards information circulating on social media compared to users that are using Twitter as usual.

All hypotheses for attitudinal effects (hypothesis 4 to hypothesis 6) are considered secondary hypotheses for the purposes of this registered report and our multiple hypotheses testing procedures. See our design table 1 summarizing the hypotheses and appropriate tests.

Table 1 Design Table

Primary Research Question	Hypothesis	Sampling Plan	Analysis Plan	Interpretation given to different outcomes
To what extent does sustained exposure to fact-checking information in a real-world environment on social media affect misinformation resilience?	H1: Exposure to fact-checking content will increase subjects' ability to accurately discern between TRUE and FALSE information compared to users that are using Twitter as usual	Our minimum sample size is 1,500 total respondents. A simulation power analysis using pilot data found that, for hypothesis 1, we are able to detect a minimum effect size of .2 standard deviation with 0.95 power using the proposed sample size	Our primary specification will use covariate-adjusted Intent-To-Treat estimates (Cov-ITT) selected via LASSO adjustment among the entire range of pre-treatment covariates. Our secondary specification will report Complier Average Causal Effect (CACE) models with the same set of pre-treatment covariates using an IV setup. See the "pre-registered analysis" section for a complete description of measurement choices and models	If we find a significant positive effect of the treatment condition, we will reject the null and find evidence for the alternative hypothesis that sustained exposure to Fact-Checking increases subjects' ability to accurately discern between TRUE and FALSE information. We will use the False Discovery Rate (FDR) method for adjusting for multiple hypotheses, considering three primary hypotheses.
	H2: Exposure to fact-checking content will increase subjects' misinformation literacy compared to users who are using Twitter as usual	Our minimum sample size is 1,500 total respondents. A simulation power analysis using pilot data found that, for hypothesis 2, we are able to detect a minimum effect size of .2 standard deviation with 0.95 power using the proposed sample size	Our primary specification will use covariate-adjusted Intent-To-Treat estimates (Cov-ITT) selected via LASSO adjustment among the entire range of pre-treatment covariates. Our secondary specification will report Complier Average Causal Effect (CACE) models with the same set of pre-treatment covariates using an IV setup. See the "pre-registered analysis" section for a complete description of measurement choices and models	If we find a significant positive effect of the treatment condition, we will reject the null and find evidence for the alternative hypothesis that sustained exposure to Fact-Checking increases subjects' misinformation literacy. We will use the False Discovery Rate (FDR) method for adjusting for multiple hypotheses, considering three primary hypotheses.
	H3:Exposure to fact-checking content will decrease subjects' conspiratorial predispositions compared to users who are using Twitter as usual	Our minimum sample size is 1,500 total respondents. A simulation power analysis using pilot data found that, for hypothesis 3, we are able to detect a minimum effect size of .2 standard deviation with 0.95 power using the proposed sample size	Our primary specification will use covariate-adjusted Intent-To-Treat estimates (Cov-ITT) selected via LASSO adjustment among the entire range of pre-treatment covariates. Our secondary specification will report Complier Average Causal Effect (CACE) models with the same set of pre-treatment covariates using an IV setup. See the "pre-registered analysis" section for a complete description of measurement choices and models.	If we find a significant positive effect of the treatment condition, we will reject the null and find evidence for the alternative hypothesis that sustained exposure to Fact-Checking decreases subjects' conspiratorial predispositions. We will use the False Discovery Rate (FDR) method for adjusting for multiple hypotheses, considering three primary hypotheses.

Design Table - Secondary Hypothesis (Continued)

Secondary Research Question	Hypothesis	Sampling Plan	Analysis Plan	Interpretation given to different outcomes
To what extent does sustained exposure to fact-checking information affect attitudes, such as trust in the media, in politicians, and information found on social media?	H4: Exposure to fact-checking content will decrease subjects' trust in the mainstream media compared to users that are using Twitter as usual..	Our minimum sample size is 1,500 total respondents. A simulation power analysis using pilot data found that, for hypothesis 4, we are able to detect a minimum effect size of .2 standard deviation with 0.95 power using the proposed sample size	Our primary specification will use covariate-adjusted Intent-To-Treat estimates (Cov-ITT) selected via LASSO adjustment among the entire range of pre-treatment covariates. Our secondary specification will report Complier Average Causal Effect (CACE) models with the same set of pre-treatment covariates using an IV setup. See the "pre-registered analysis" section for a complete description of measurement choices and models.	If we find a significant positive effect of the treatment condition, we will reject the null and find evidence for the alternative hypothesis that sustained exposure to Fact-Checking decreases subjects' trust in the mainstream media. We will use the False Discovery Rate (FDR) method for adjusting for multiple hypotheses for both primary and secondary hypotheses, thus correcting for six hypotheses.
	H5: Exposure to fact-checking content will increase subjects' cynicism towards politics compared to users that are using Twitter as usual.	Our minimum sample size is 1,500 total respondents. A simulation power analysis using pilot data found that, for hypothesis 5, we are able to detect a minimum effect size of .2 standard deviation with 0.95 power using the proposed sample size	Our primary specification will use covariate-adjusted Intent-To-Treat estimates (Cov-ITT) selected via LASSO adjustment among the entire range of pre-treatment covariates. Our secondary specification will report Complier Average Causal Effect (CACE) models with the same set of pre-treatment covariates using an IV setup. See the "pre-registered analysis" section for a complete description of measurement choices and models.	If we find a significant positive effect of the treatment condition, we will reject the null and find evidence for the alternative hypothesis that sustained exposure to Fact-Checking increases subjects' political cynicism. We will use the False Discovery Rate (FDR) method for adjusting for multiple hypotheses, for both primary and secondary hypotheses, thus correcting for six hypotheses.
	H6: Exposure to fact-checking content will increase subjects' cynicism towards information circulating on social media compared to users that are using Twitter as usual.	Our minimum sample size is 1,500 total respondents. A simulation power analysis using pilot data found that, for hypothesis 6, we are able to detect a minimum effect size of .2 standard deviation with 0.95 power using the proposed sample size	Our primary specification will use covariate-adjusted Intent-To-Treat estimates (Cov-ITT) selected via LASSO adjustment among the entire range of pre-treatment covariates. Our secondary specification will report Complier Average Causal Effect (CACE) models with the same set of pre-treatment covariates using an IV setup. See the "pre-registered analysis" section for a complete description of measurement choices and models.	If we find a significant positive effect of the treatment condition, we will reject the null and find evidence for the alternative hypothesis that sustained exposure to Fact-Checking increases subjects' social media cynicism. We will use the False Discovery Rate (FDR) method for adjusting for multiple hypotheses, for both primary and secondary hypotheses, thus correcting for six hypotheses.

Methods

Ethics information

This research proposal, including the procedures for collecting pilot data described in this proposal, received approval from the Institutional Review Board of New York University (IRB-FY2023-6870). All the study protocols were revised by the IRB, including the consent form of the surveys and information about participants' compensation.

Pilot data and Sampling Plan

Between May 21 and June 22, 2023, we ran a pilot study that followed the design described in the previous section, but with fewer participants. We recruited 114 participants for the pilot study through Facebook Ads, and from this sample, 104 participants completed the final post-treatment survey (completion rate of 91%). We use the pilot data to a) estimate a solid-grounded power analysis using information from the pilot data and b) present our pre-registered models and measurement choice as we commit to do in the full manuscript.

First, we utilize our collected pilot study data for all of our dependent variables to plot the power for an unadjusted Intent-to-Treat ITT model for sample sizes from 50 to 2,000 for all six of our hypotheses. Power is calculated using *DeclareDesign* (45) and randomly sampling from the control group data collected during the pilot and creating treatment group data assuming a population average treatment effect (PATE) of 0.2 Cohen's d, which is often described as the threshold for a negligible effect (46). Figure 2 shows that assuming a PATE of 0.2 Cohen's d, a sample size of 1,500 in the final survey should report a statistically significant effect size (P-value < 0.05) over 95% of the time using an unadjusted ITT model for all of our hypotheses. Assuming levels of attrition higher than those described for the pilot, we propose to recruit

1,700 participants for the pre-treatment survey.

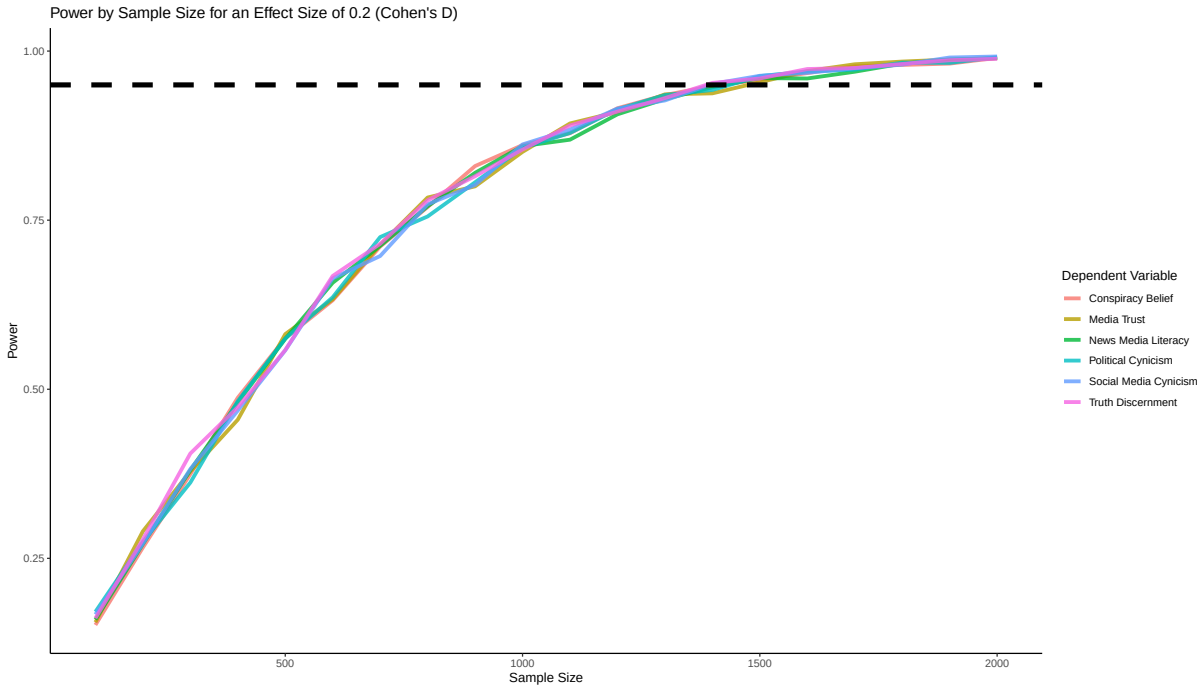


Figure 2 Power Analysis

Design

We will conduct an online field experiment in which we substantially increase social media users' exposure to fact-checking accounts and the content produced by these organizations in social media users' own news feeds. To this end, we randomly assign Twitter users to an intervention that creates a new timeline at the top of participants' Twitter feeds. This timeline is composed of a pre-curated list of active fact-checking organizations on Twitter. This new timeline will appear on the side of any other timeline users have on their Twitter home page. We will not blind participants to their group allocation.

Figure 3 presents a stylized description of the experimental design, and we discuss the different stages of the design in further detail below.

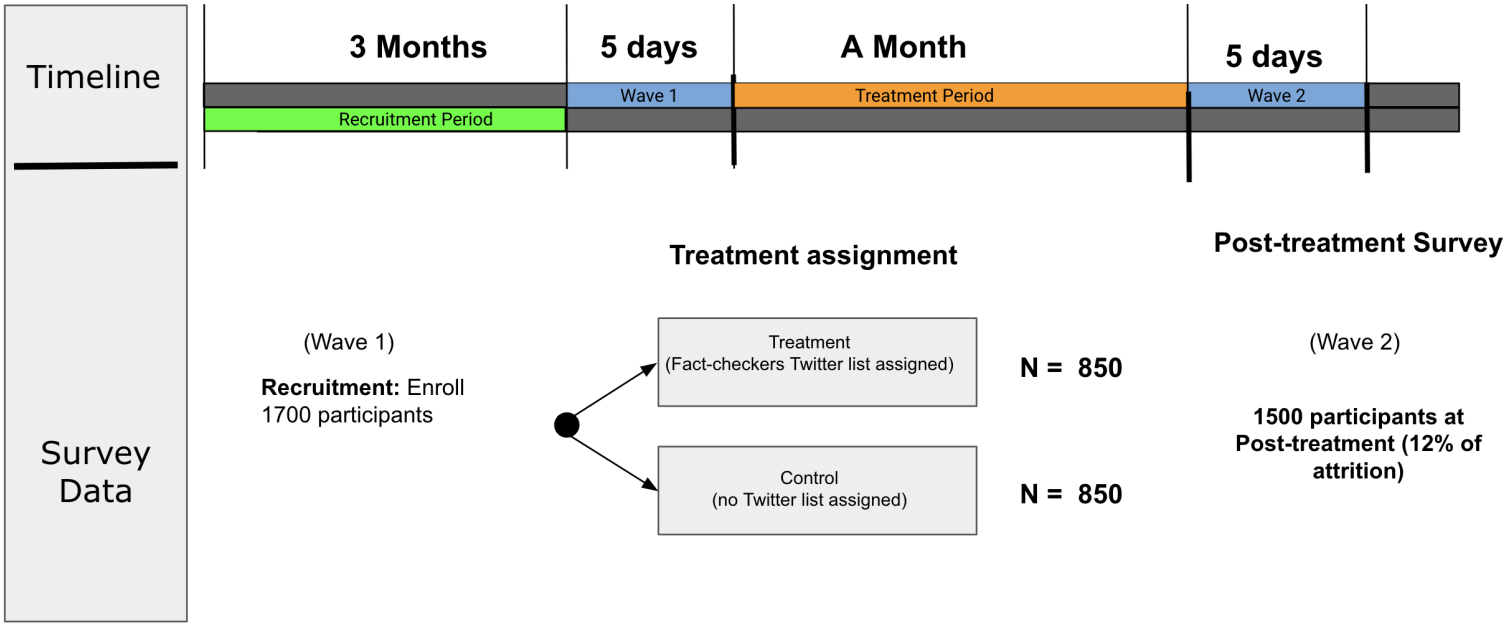


Figure 3 Overview of the Fact-Checking Experiment

Stage 1 - Recruitment: Participants will be recruited using Online Ads (Facebook and Twitter). At this first stage, participants will answer a quick recruitment survey. This survey will contain a short battery of questions about respondents' demographics, their social media habits, particularly about Twitter, their willingness to join our study, and their authorization for our research team to create the Twitter Timeline with Fact-Checking accounts on their Twitter account if participants are assigned to this condition. Appendix E presents the recruitment survey.

To create the Twitter Timeline with fact-checking organizations, we develop a Twitter application that collects subjects' permission for our research team to pin a Twitter list to their home timeline. Authorization for the app occurs at the end of the recruitment survey, and we request all the participants to provide us with access. After receiving the user authorization, the process for creating the new timeline requires only a simple post request to the Twitter API. In addition to the programmatic action to create the Timeline, we will also provide participants with a manual tutorial on how to join the treatment group. This allows us to overcome technical difficulties from the Twitter API that might affect treatment take-up and compliance. Compared to the pilot we present in this Registered Report, allowing users also to join the experiment via a manual setup substantively increases the feasibility of our study since we will consider now eligible to join the study both participants who authorize our access to their Twitter API and participants who agreed to join the study but failed to give us access to the APIs.

Based on the pre-treatment survey, we will use five variables to select the participants who will be invited to join the Twitter Fact checking experiment:

- Participants should be residing in the United States
- Participants should be older than 18
- Participants should agree to participate in the experiment (last question in the survey)

- Participants should use Twitter for more than 10 min per day

We will also remove low-quality responses, including participants who speed through the recruitment survey, participants from the same IP address, and participants classified as bots or automatic accounts by Qualtrics filters.

Stage 2 - Treatment Intervention and Pre-Treatment Outcomes: After recruiting the number of participants necessary to start the study given our power analysis, we will invite participants by email to join the full study. In the email, participants will be invited to take a survey to collect pre-treatment outcomes. At the end of this pre-treatment survey, we will present to participants their treatment assignment condition and confirm their intentions to participate in the study.[3] Appendix E presents the pre-treatment survey.

Participants will be randomized into two distinct groups:

- Treatment: Subjects will have added a new timeline to their Twitter feed with a list of pre-curated Fact-checking media organizations and will be incentivized to look at the new timeline.

- Control: The control group will not be asked to change anything in their Twitter feed.

The pre-curated fact-checking organizations were selected based on a list by the Duke Reporters' Lab, which lists 74 active fact-checking organizations.[4] From this list, we selected accounts that (1) have a Twitter account, (2) appear to have a neutral partisan bias, (3) where 95% of tweets are fact-checks, and (4) cover the whole United States (i.e. no international focus, no regional focus). A total of eight accounts satisfied all four criteria.[5]

[3]The placement of the treatment assignment at the end of the pre-treatment survey is a deviation of the pilot. As we discussed further in the analysis section, presenting the treatment assignment before collecting pre-treatment outcomes might generate post-treatment bias, just as a priming effect of informing participants about their commitment to consuming fact-checking content. Therefore, we decided to adjust this design choice for the full study.

[4]https://reporterslab.org/fact-checking

[5]Selected organizations: @VerifyThis, @APFactCheck, @LeadStoriesCom, @APFactCheck, @erumors, @factcheckdotorg, @snopes, @PolitiFact

Participants will be split evenly between the two experimental groups. We will block randomize the treatment assignment using age, education, partisanship, and gender.[6]

Stage 3 - Compliance Tasks: In addition to facilitating users' exposure to fact-checking organizations on social media, our ability to make inferences from our study requires ensuring that participants are indeed consuming the content from their new Timeline. Therefore, we devise a set of incentives to increase treatment take-up in the field experiment.

To increase compliance with the intervention, we will incentivize participants to scroll down through their fact-checking timeline. To this end, we will offer participants the opportunity to increase their financial compensation by sending us daily screenshots of the tweets from the fact-checking organizations added from the timeline.[7] More specifically, we will ask them to send us a tweet from this new timeline with fact-checking accounts that they found most informative that day. To ensure balance in time on Twitter across the control and treatment group, we will also offer similar incentives to the control group to send us daily screenshots of a tweet they found most informative on their Twitter feed on any given day.[8] In this way, we guarantee that treatment and control participants receive comparable incentives to regularly check their Twitter feeds and, therefore, that any effects we pick up are not simply the result of additional attention or time spent on Twitter due to the need to select tweets for the screenshots.

Requests to complete the compliance tasks will be sent daily to all participants and will be collected through Qualtrics surveys. Section E of the Appendix, Compliance Survey, presents the full description of the task.

[6]Participants in the treatment and control group will receive $2 dollars for completing this survey.

[7]If participants send us at least 10 screenshots on different days, we offer an additional $18, between 5 and 10 screenshots on different days, we offer an additional $13, and not extra compensation if they send less than 5 screenshots on different days

[8]If participants send us at least 10 screenshots on different days, we offer an additional $8, between 5 and 10 screenshots on different days, we offer an additional $3, and not extra compensation if they send less than 5 screenshots on different days. These incentives are different than the structure offered to the treated participants because we assume participants in the treatment will spend more time and effort on the newly created timeline than control participants just looking through their regular timeline.

Stage 4 - Post-Treatment Survey: Four weeks after the beginning of the treatment assignment, participants will be invited to answer a post-treatment survey in which our team will measure the pre-registered outcomes described in the following sections. Appendix E, section Post-Treatment Survey, presents the full questionnaire for this final survey.[9]

Analysis plan

In this section, we present and discuss the pre-registered models we intend to report in the full manuscript. The purpose of the sections is to provide full transparency for our analysis pipeline, including all decisions related to measurement and modeling choices. We use the raw data collected during the pilot study explained earlier to describe our pipeline. Our pre-registered models use results from two different specifications. Our primary specification will use covariate-adjusted Intent-To-Treat estimates (Cov-ITT) selected via LASSO adjustment among the entire range of pre-treatment covariates collected in the recruitment and pre-treatment survey (See appendix E, sections recruitment survey and pre-treatment survey). We include both pre-treatment covariates and pre-treatment outcomes as eligible controls for the LASSO covariate selection. Our secondary specification will report Complier Average Causal Effect (CACE) models with the same set of pre-treatment covariates selected via LASSO adjustment using an IV setup, instrumenting the compliance measure with an indicator for treatment assignment.[10] For all pre-registered models, we present 95% robust confidence intervals.

If the registered report is accepted for publication, then we also commit to reporting the

[9] We offer participants a $5 dollar compensation for completing the final survey

[10] In our pilot, we first informed the participants about their treatment assignment and later measured pre-treatment outcomes. Even though our treatment (in the field consumption of fact-checking content) only occurs after responses are collected, we are aware that post-treatment bias might emerge from the revelation of the experimental conditions. Therefore, to be conservative with the estimation of the effect sizes, we use a restricted set of pre-treatment covariates, removing all pre-treatment outcomes and using as controls only pre-treatment covariates that we do not predict to be affected by the fact-checking intervention. For the full study, we will only reveal the treatment condition at the end of the survey and, therefore, use all the pre-treatment outcomes and covariates in the LASSO adjustment selection.

intention-to-treat effects without any covariate adjustment – which are omitted from the next sections due to the limits of the sample size collected in the pilot – in the final published version of the paper. We will use HC2 robust standard errors in all analyses and report p-values from two-tailed t-tests. As a robustness test, we will adjust for multiple hypothesis testing considering three hypotheses for misinformation outcomes (h1, h2, and h3), and all seven hypotheses for the attitudinal effects (secondary hypotheses) on trust, and cynicism. We will use the False Discovery Rate (FDR) method for adjusting for multiple hypotheses testing. Complier Average Causal Effect (CACE) models will be estimated as described using the pilot data using an IV setup. Finally, for the CACE models, we will consider participants who sent at least ten screenshots throughout the four weeks to have complied. We also commit to providing results with alternative measures of compliance.

In the appendix, we present the same results using a simulated sample of 1,800 observations using resampling methods and repeat this with a bootstrapping estimate 1,000 times to build credible intervals for the predicted effects of our specified models. The point estimates across both the raw and the synthetic data are, as expected, very close. However, the synthetic data provides the reader with an approximation of what our confidence intervals are most likely to be in the full manuscript.

Treatment Take-up: We start by documenting substantial and sustained levels of treatment take-up during the one-month pilot study. During the study, we incentivized participants to consume the content presented in the *Fact-Checking Timeline*[11] by asking them to send us screenshots or links from posts that appeared in the timeline that participants considered most informative. Meanwhile, to avoid a bundled treatment, participants in the control group were asked to do the same tasks but solely using their regular Twitter timelines.

We document results from these tasks in figure 4. The left panel presents the full distribution

[11] As described before, for the stage 2 study, we will relabel the timeline as *Media Timeline*, to avoid demand effects in the treatment group. Results from the pilot use the label *Fact-Checking Timeline*

of the sum of tasks completed by every participant in the treatment and control groups. Overall, subjects' participation in these tasks was considerably high, indicating a successful implementation of the proposed design. The median participant completed the tasks twenty times over the one-month treatment (mean=17.69, SD=7.4). Out of 104 participants who completed the study, only 16 completed the tasks less than ten times (84% of compliance). Using a Komolgorov test to compare the distributions between the number of completed tasks per participant among the groups, we cannot reject the null that these densities emerge from a similar distribution ($p - value$ = 0.3803).

The right panel details the analyses showing the density for the sum of tasks in which participants successfully submitted a tweet from accounts that are part of the *Fact-Checking Timeline*. To measure success, we classify the author of every tweet submitted by the participants (in the treatment) via screenshots or links. We verified whether these messages were indeed coming from Fact-Checking Organizations included in the treatment intervention. The median treated participant submitted 19 screenshots/links from fact-checking organizations during the one-month period, close to one link per day, and only 11 out of 51 participants in the treatment group completed fewer than 10 tasks correctly (82% of compliance), indicating that the vast majority of participants in the treatment indeed sent us information from the fact-checking organizations added to the new timeline. These numbers still suggest a robust, sustained exposure to fact-checking content over the treatment period.

This section presents the pre-registered results precisely, as we intend to report in the full manuscript. All the figures in these sections should be interpreted as empty shells serving the purpose of providing transparency for our entire analysis pipeline, including all decisions related to measurement and modeling choices. To populate the figures, we use the raw data collected during the pilot described before. Given the relatively small sample size collected for the pilot, the standard deviations of the models are considerably wide. We refrain from providing

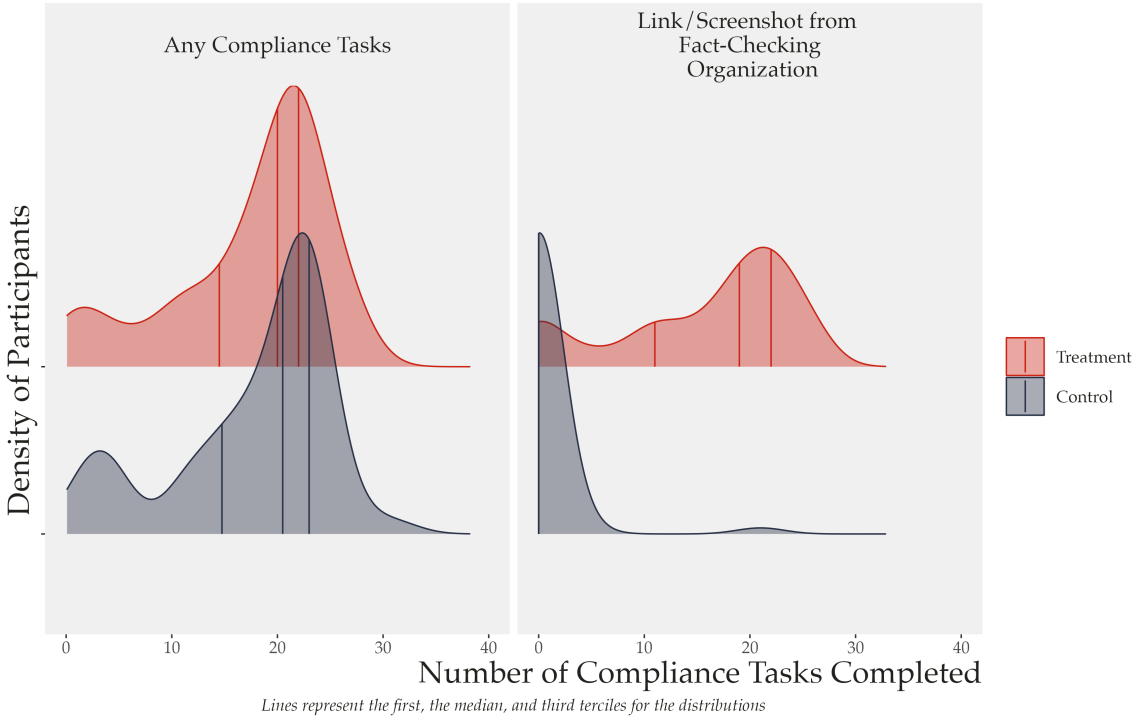


Figure 4 Pilot Results: Treatment Take-up

any interpretation of the statistical significance of the results and the point estimates. The pilot data here serves the purposes of assessing the feasibility of the experiment and estimating the effect size and variance of the outcomes to inform our power analysis. These elements are taken into consideration in the power analysis reported in the previous section.

Primary Hypotheses: Effects of Sustained Exposure to Fact-Checking on Resilience to Misinformation

Truth Discernment: Our first hypothesis expects that sustained exposure to fact-checking information will increase subjects' ability to discern TRUE news and FALSE rumors information. To measure discernment, we will ask participants to rate the accuracy of ten headlines, five of which are true news and five are false rumors. The true headlines will be selected from mainstream news outlets, and the false headlines will be selected from corrected published fact-checks posted by the fact-checking organizations included in the Fact-Checking Timeline during the treatment period. All respondents will be asked to evaluate the accuracy of the article using a 4-point scale ("Not at all Accurate" to "Very Accurate").

Using responses from the headline tasks, we pre-register three different outcomes: *False Rumors Accuracy*, *True News Accuracy*, and *Truth Discernment*. *False Rumors Accuracy* counts the number of false headlines classified by respondents as "Not at all accurate" or "Not very accurate"; meanwhile, *True News Accuracy* counts the number of true headlines judged by respondents as "Somewhat accurate" or "Very accurate". Since these are not force-based tasks, we will recode *Don't Know* responses as zero in the sum for each outcome. *Truth Discernment* represents a composite measure summing both *False Rumors Accuracy* and *True News Accuracy* (see table 2 in the supplemental materials), in other words, taking into consideration the participants' capacity to discern between true and false headlines. In addition to estimating models using the sum of the accurate responses, following (47) suggestion, we will also present item-level models using random intercepts to control for headlines and individual-level

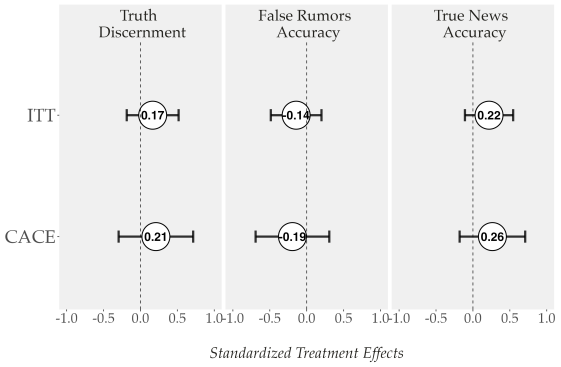
variation. Figure 5A presents our pre-registered analysis as we intend to report in the main manuscript using the raw data from the pilot design.

Conspiratorial predispositions: In addition to measuring the treatment effects on the capacity to discern true and false information using recently published headlines, as an additional outcome, we will measure treatment effects on long-standing conspiratorial predispositions. We will use a set of four items (48) and ask participants their agreement with each using a 5-point agreement scale. We will standardize responses from these four measures and use the sum of their z-scores to build a composite index (*conspiratorial predispositions*) as our primary outcome of interest. Higher values in the composite index indicate stronger conspiratorial predispositions. Figure 5B presents our pre-registered analysis as we intend to report in the main manuscript using the raw data from the pilot previously discussed.

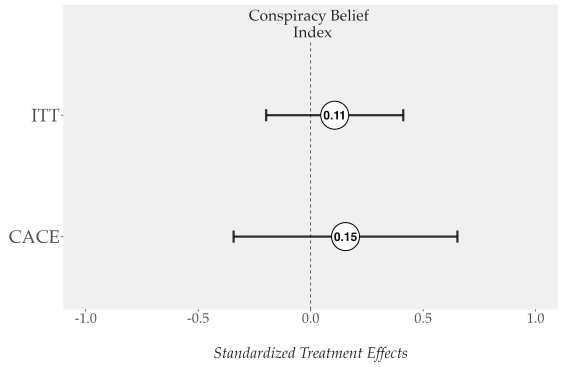
Misinformation Literacy: Hypothesis 3 expects that sustained exposure to fact-checking correction will also promote long-term changes in subjects' misinformation literacy. To capture this effect, we will use a misinformation literacy task, which consists of seven items about participants' habits in accessing, sharing, and evaluating the credibility of information consumed on social media. As before, we will build a standardized measure using these items (*Misinformation Literacy*) in which higher values mean greater literacy in evaluating misinformation on online news. Figure 5C presents our pre-registered analysis as we intend to report in the main manuscript using the raw data from the pilot previously discussed.

Secondary Hypotheses: Effects of Sustained Exposure to Fact-Checking on Attitudinal Effects

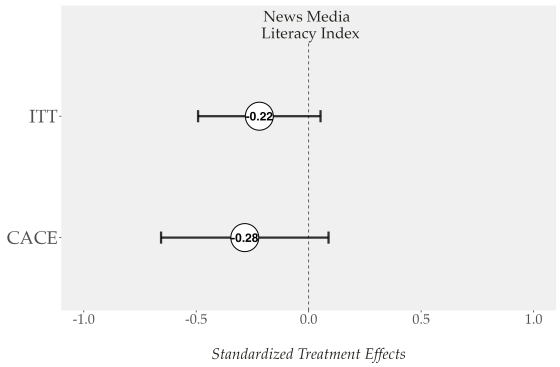
Trust in the Media: To measure trust in the media (*hypothesis 4*), we will build an index (*Media Trust Scores*) summing subjects' responses to three items asking about trust in news mainstream media, balance in news covered by these organizations, how often the mainstream media fabricate news. We will convert responses to a numerical scale and sum their recoded



(a) Truth Discernment



(b) Conspiratorial Predispositions



(c) Misinformation Literacy

Figure 5 Pilot Results on the misinformation resilience outcomes. Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% confidence intervals

values to build our outcome of interest. Figure 6A presents our pre-registered analysis as we intend to report in the main manuscript using the raw data from the pilot previously discussed.

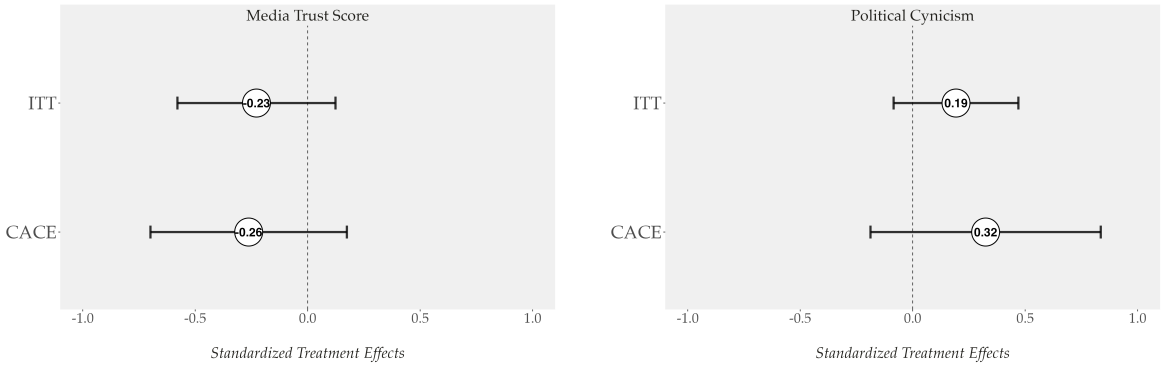
Political Cynicism: To measure the attitudinal effects of sustained exposure to fact-checking on skepticism about politics (*hypothesis 5*), we will use two survey questions measuring trust in democracy and trust in politicians in the United States. We will build a *Political Cynicism Index* using these two outcomes, which consists of the sum for all respondents of the standardized response (z-score) across the three social media cynicism questions. Figure 6B presents our pre-registered analysis as we intend to report in the main manuscript using the raw data from the pilot previously discussed.

Social Media Cynicism: To measure the attitudinal effects of sustained exposure to fact-checking on skepticism of information consumed through social media (*hypothesis 6*), we will use three survey questions measuring trust in news stories shared on social media, the prevalence of misinformation on social media news, and the use of social media for news information. We build a *Social Media Cynicism Index* using these three outcomes, which consists of the sum for all respondents of the standardized response (z-score) across the three social media cynicism questions. Figure 6C presents our pre-registered analysis as we intend to report in the main manuscript using the raw data from the pilot previously discussed.

Additional Analysis

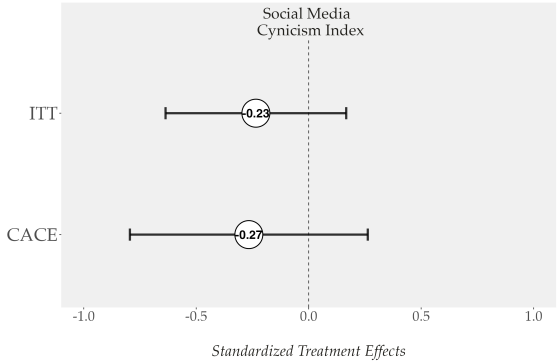
In this section, we describe a set of exploratory, non-causal additional analyses we intend to report as additional analyses in the manuscript. These analyses explore additional outcomes and different measurement choices and provide robustness for the primary analysis of the manuscript.

Affective Polarization: Apart from the effects of fact-checking on misinformation resilience, trust in the media and cynicism, we are also interested in the effects of fact-checking



(a) Media Trust

(b) Political Cynicism



(c) Social Media Cynicism

Figure 6 Pilot Results on Attitudinal Effects. Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% confidence intervals

information on attitudes towards political groups and institutions in society. Past work has shown that fact-checking can reduce belief in falsely held beliefs about political or ethnic outgroups caused by exposure to misleading content (49, 50, 51, 52). Correcting these beliefs could change attitudes towards outgroups or even innoculate them from rhetoric that paints a political or ethnic out-group in a negative manner. Therefore, as an additional research question, we ask: To what extent does sustained exposure to fact-checks on social media affect attitudes toward outgroup voters? Results for this research question will be reported as additional, exploratory, and non-causal findings. Power analysis from our pilot data indicated that this particular outcome requires a much larger sample size to achieve 95% power, holding constant effect sizes. The reason comes from the fact that these variables exhibit low levels of variance in our sample and ceiling effects with many participants being highly polarized. Therefore, without the appropriate statistical power, we decided to incorporate this outcome as an additional exploratory analysis in the manuscript.

To measure the treatment effects on outgroup polarization, we will use four different outcomes: *Party Affective Polarization*, *Candidate Affective Polarization*, *Outgroup Feelings*, and *Ingroup Feelings*. To build these outcomes, we will use a set of well-established feeling thermometer asking how participants feel about the two major parties in the US and their last presidential candidates. The *Party Affective Polarization* takes the absolute difference in the feeling thermother for Democrats and Republicans, the *Candidate Affective Polarization* uses the same measure but regards Joe Bidden and Donald Trump. Meanwhile, *Outgroup Feelings* and *Ingroup Feelings* take the self-reported feeling towards Democrats/Republicans for each voter conditional on their self-reported partisanship. Figure 19 in Appendix F presents our results as we intend to report in the additional analysis section of the main manuscript using the simulated data from the pilot previously discussed.

Perceived Ability and Overconfidence on Misinformation Judgement: In addition to us-

ing the headlines task to measure news veracity discernment, we will also measure the effects of sustained exposure to fact-checking information on respondents' self-reported ability to identify false information online. Furthermore, we will combine both measures - *False Rumors Accuracy* from the headlines task and self-reported ability - to build a measure for overconfidence measure on respondents' misinformation judgment (53). Since this outcome has not been widely tested in the literature, we present here as an additional analysis, and do not include it in our pre-registered hypotheses. However, we intend to present it as an explanatory analysis in the main manuscript. Figure 23 in Appendix F presents our pre-registered analysis as we intend to report in the main manuscript using the simulated data from the pilot previously discussed.

Decomposing Direct and Indirect Treatment Effects: Our quantity of interest in the Registered Report is the total average treatment effects of our intervention. However, as described in the causal inference literature (54, 55, 56), total effects can be broadly decomposed into direct effects of the intervention on the outcome, and indirect effects, which are mediated by a third variable. Although we do not pre-register any mediation analysis from specific covariates, an important concern is the possibility of a common underlying factor driving our findings across our six pre-registered outcomes.[12] If the outcomes are strongly correlated, then we should consider the possibility that indirect effects dominate most of the total effects identified by our models. To address this concern, we commit to report in the supplementary materials a full mediation analysis decomposing indirect and direct effects for all the pre-registered outcomes, simultaneously using all the other pre-registered outcomes as mediators. Interpreting our direct estimates as causal requires sequential ignorability and the independence of the mediators (54, 55). For this reason, our mediation models (1) control for all LASSO-selected pre-treatment covariates and (2) include simultaneously all outcomes available for indirect effects in the second-stage analysis.

[12] We thank Reviewer 2 for raising this issue

Data availability

The data and code to reproduce all the analysis will be made available in a public repository upon publication of the research. The pilot data and the code to reproduce the analysis presented in the stage I manuscript are available on this GitHub repository (https://github.com/SMAPPNYU/twitter-lists-experiment-final) for peer review.

Code availability

The data and code to reproduce all the analyses will be made available in a public repository upon publication of the research. The pilot data and the code to reproduce the analysis presented in the stage I manuscript are available for peer review on this GitHub repository (https://github.com/SMAPPNYU/twitter-lists-experiment-final).

Acknowledgements

We are grateful to all New York University (NYU) Center for Social Media and Politics members for their excellent comments and suggestions. We thank Leticia Bode for helpful feedback; Angie Waller, Sarah Graham and Edwin Kamau for their support in managing the survey panel and developing the app; and Maitreyi Natarajan for the excellent research assistance.

Author contributions

Conceptualization: TV, KA and JT; Methodology: TV, KA, FL, JT and SM; Investigation: TV, KA, FL; Visualization: TV and KA; Supervision: SM; Writing—original draft: TV, KA, FL; Writing—review editing: TV, KA, FL, JT and SM

Competing interests

JT received a one-time fee from Facebook, the parent company of WhatsApp, to compensate him for administrative time spent in organizing a 1-day conference for approximately 30 academic researchers and a dozen Facebook product managers and data scientists that was held at NYU in the summer of 2017 to discuss research related to civic engagement; the fee was paid before any work or data collection for the current project began. He did not provide any consulting services nor any advice to Facebook as part of this arrangement. JT is currently a co-chair of the external academic team for the U.S. 2020 Facebook & Instagram Election Study, a research collaboration between a team of external academic researchers and internal Meta researchers; he receives no financial compensation from Meta for this work. SM is a former employee of Meta; his employment concluded before any work on this project began. JT is currently a Senior Geopolitical Risk Advisor at Kroll. The authors declare no conflicts of interest.

References

[1] Bovet A, Makse HA (2019) Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10:7.

[2] Recuero R, Soares FB, Gruzd A (2020) Hyperpartisanship, Disinformation and Political Conversations on Twitter: The Brazilian Presidential Election of 2018. *Proceedings of the International AAAI Conference on Web and Social Media* 14:569–578.

[3] Mello PC (2020) *A máquina do ódio: notas de uma repórter sobre fake news e violência digital*. (Companhia das Letras).

[4] Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ (2021) Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour* 5(3):337–348.

[5] Van der Linden S, Leiserowitz A, Rosenthal S, Maibach E (2017) Inoculating the public against misinformation about climate change. *Global challenges* 1(2):1600008.

[6] Saha P, Mathew B, Garimella K, Mukherjee A (2021) "short is the road that leads from fear to hate": Fear speech in indian whatsapp groups in *Proceedings of the Web Conference 2021*. pp. 1110–1121.

[7] Banaji S, Bhat R, Agarwal A, Passanha N, Sadhana Pravin M (2019) Whatsapp vigilantes: An exploration of citizen reception and circulation of whatsapp misinformation linked to mob violence in india. *Working Paper*.

[8] Walter N, Cohen J, Holbert RL, Morag Y (2020) Fact-checking: A meta-analysis of what works and for whom. *Political Communication* 37(3):350–375.

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

[10] Porter E, Wood TJ (2021) The global effectiveness of fact-checking: Evidence from simultaneous experiments in argentina, nigeria, south africa, and the united kingdom. *Proceedings of the National Academy of Sciences* 118(37):e2104235118.

[11] Brashier NM, Pennycook G, Berinsky AJ, Rand DG (2021) Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences* 118(5):e2020043118.

[12] Bode L, Vraga EK (2018) See something, say something: Correction of global health misinformation on social media. *Health communication* 33(9):1131–1140.

[13] Blair RA, et al. (2023) Interventions to counter misinformation: lessons from the global north and applications to the global south. *Current Opinion in Psychology* p. 101732.

[14] Guess AM, et al. (2023) How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science (New York, N.Y.)* 381:398–404.

[15] Guess AM, et al. (2023) Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science (New York, N.Y.)* 381:404–408.

[16] Wojcik S, et al. (2022) Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation.

[17] Allen J, Martel C, Rand DG (2022) Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program in *CHI Conference on Human Factors in Computing Systems*. (ACM), pp. 1–19.

[18] Arceneaux K, Johnson M, Murphy C (2012) Polarized political communication, oppositional media hostility, and selective exposure. *The Journal of Politics* 74(1):174–186.

[19] Bail CA, et al. (2018) Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115(37):9216–9221.

[20] Bowles J, Croke K, Larreguy H, Liu S, Marshall J (2022) Sustained exposure to fact-checks can innoculate citizens against misinformation in the global south. *URL: https://www. dropbox. com/s/a6rnh3o7c97dcqn/WCW_Science_Submission. pdf*.

[21] Broockman D, Kalla J (2022) The impacts of selective partisan media exposure: A field experiment with fox news viewers. *OSF Preprints* 1.

[22] Pennycook G, Rand DG (2019) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188:39–50.

[23] Pennycook G, et al. (2021) Shifting attention to accuracy can reduce misinformation online. *Nature* 592(7855):590–595.

[24] Kunda Z (1990) The case for motivated reasoning. *Psychological bulletin* 108(3):480.

[25] Taber CS, Lodge M (2006) Motivated skepticism in the evaluation of political beliefs. *American journal of political science* 50(3):755–769.

[26] Carey JM, et al. (2022) The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature Human Behaviour* 6:236–243.

[27] Fazio LK, Rand DG, Pennycook G (2019) Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic bulletin & review* 26:1705–1710.

[28] Arendt F (2015) Toward a dose-response account of media priming. *Communication Research* 42(8):1089–1115.

[29] Roozenbeek J, Van Der Linden S, Goldberg B, Rathje S, Lewandowsky S (2022) Psychological inoculation improves resilience against misinformation on social media. *Science advances* 8(34):eabo6254.

[30] Roozenbeek J, Van der Linden S (2019) Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5(1):1–10.

[31] Pereira FB, Bueno NS, Nunes F, Pavão N (2022) Inoculation reduces misinformation: Experimental evidence from multidimensional interventions in brazil. *Journal of Experimental Political Science* pp. 1–12.

[32] Guess AM, et al. (2020) A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117:15536–15545.

[33] Aslett K, Guess AM, Bonneau R, Nagler J, Tucker JA (2022) News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances* 8:eabl3844.

[34] Marques FPJ, Ferracioli P, Comel N, Kniess AB (2023) Who is who in fact-checked conspiracy theories? Disseminators, sources, and the struggle for authority in polarized environments. *Journalism* p. 146488492311655.

[35] Harambam J (2021) Conspiracy theories: Misinformed publics or wittingly believing false information? in *The Routledge Companion to Media Disinformation and Populism*. (Routledge), pp. 302–311.

[36] Kreko P (2020) Countering conspiracy theories and misinformation in *Routledge handbook of conspiracy theories*. (Routledge), pp. 242–256.

[37] Cappella JN, Jamieson KH (1997) *Spiral of cynicism: The press and the public good*. (Oxford University Press).

[38] Dias N, Sippitt A (2020) Researching Fact Checking: Present Limitations and Future Opportunities. *The Political Quarterly* 91:605–613.

[39] Bachmann I, Valenzuela S (2023) Studying the Downstream Effects of Fact-Checking on Social Media: Experiments on Correction Formats, Belief Accuracy, and Media Trust. *Social Media + Society* 9:20563051231179694.

[40] Van Duyn E, Collier J (2019) Priming and Fake News: The Effects of Elite Discourse on Evaluations of News Media. *Mass Communication and Society* 22:29–48.

[41] York C, et al. (2020) Effects of Fact-Checking Political Misinformation on Perceptual Accuracy and Epistemic Political Efficacy. *Journalism & Mass Communication Quarterly* 97:958–980.

[42] Lee S, Jones-Jang SM (2022) Cynical nonpartisans: The role of misinformation in political cynicism during the 2020 us presidential election. *new media & society* p. 14614448221116036.

[43] Luo M, Hancock JT, Markowitz DM (2022) Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research* 49(2):171–195.

[44] Clayton K, et al. (2020) Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political behavior* 42:1073–1095.

[45] Blair G, Cooper J, Coppock A, Humphreys M (2019) Declaring and diagnosing research designs. *American Political Science Review* 113(3):838–859.

[46] Cohen J (1969) *Statistical power analysis for the behavioral sciences*. (Lawrence Erlbaum Associates).

[47] Pennycook G, Binnendyk J, Newton C, Rand DG (2021) A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology* 7(1):25293.

[48] Uscinski JE, Klofstad C, Atkinson MD (2016) What drives conspiratorial beliefs? the role of informational cues and predispositions. *Political Research Quarterly* 69(1):57–71.

[49] Druckman JN, et al. (2023) Correcting misperceptions of out-partisans decreases american legislators' support for undemocratic practices. *Proceedings of the National Academy of Sciences* 120(23):e2301836120.

[50] Voelkel JG, et al. (2023) Megastudy identifying effective interventions to strengthen americans' democratic attitudes.

[51] Hameleers M, Van Der Meer TGLA (2020) Misinformation and Polarization in a High-Choice Media Environment: How Effective Are Political Fact-Checkers? *Communication Research* 47:227–250.

[52] Fridkin K, Kenney PJ, Wintersieck A (2015) Liar, Liar, Pants on Fire: How Fact-Checking Influences Citizens' Reactions to Negative Advertising. *Political Communication* 32:127–151.

[53] Lyons BA, Montgomery JM, Guess AM, Nyhan B, Reifler J (2021) Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences* 118(23):e2019527118.

[54] Imai K, Keele L, Tingley D, Yamamoto T (2011) Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105(4):765–789.

[55] Heckman JJ, Pinto R (2015) Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric reviews* 34(1-2):6–31.

[56] Pearl J (2022) Direct and indirect effects in *Probabilistic and causal inference: the works of Judea Pearl*. pp. 373–392.

[57] Newman N, Fletcher R, Kalogeropoulos A, Nielsen RK (2019) Reuters digital news report 2019.

[58] Vraga EK, Tully M (2021) News literacy, social media behaviors, and skepticism toward information on social media. *Information, Communication & Society* 24(2):150–166.

[59] Ahmed S (2021) Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *new media & society* p. 14614448211019198.

Appendix A: Outcomes of Interest

This section describes our outcomes of interest and pre-registered measurement choices for each outcome. Covariates are fully described in the survey materials, and we will use LASSO regularization to adjudicate which pre-treatment variables will be added to the models. Table 2 summarizes our outcomes.

Table 2 Pre-Registered Outcomes and Measurement Choices

Variable Name	Definition	Measurement	Range
False Rumors Accuracy	Measures the ability to discern false information using a headline task	Sum of the true headlines classified as accurate.	0-5
True News Accuracy	Measures the ability to discern true information using a headline task	Sum of the false headlines classified as inaccurate.	0-5
Truth discernment	Measures the ability to detect false and true information using a headline task	Sum of False Rumors and True News accuracy outcomes.	0-10
Conspiratorial predispositions score	Measures subjects' conspiratorial predispositions	Sum of standardized z-score across four survey questions.	-8 to +8
Misinformation literacy score	Measures subjects' literacy in evaluating false information online	Sum of standardized z-score across six literacy items.	-12 to +12
Affective Polarization	Measures affective polarization towards the two major US parties	Absolute value of the difference between the feeling thermometers of each political party	0-100.0
Media Trust Score	Measures the respondent's trust in media with three separate question about participants trust in information from mainstream media.	To create an index for media trust we sum the values assigned to each respondent's answer for each question.	0.0-3.0
Political cynicism	Measures the respondent's level of two items for political cynicism	Uses the z-score sum of two survey items.	-6 to 6
Social Media Cynicism	Measures how skeptical the respondent is towards getting news on social media	Uses the z-score sum of three survey items	-9 to 9
Perceived Ability in Misinformation Judgment	Measures the perceived ability to detect misinformation	Average value of two self-assessments about the ability to detect misinformation (0.0=no perceived ability, 100.0=high perceived ability).	0.0-100.0

We proceed now with a full description of the outcomes, including the wording for all survey outcomes.

- **Headlines Task**: Ability to detect misinformation and true information (headline task): To construct this measure, we ask respondents to evaluate the accuracy of 10 headlines on a 4-point scale ranging from very accurate (4) to not at all accurate (1). Actual news sources published all of the headlines within one month of the post-treatment survey, and a portion of the headlines was rated as false by at least one third-party fact-checking organization. The order of the headlines was randomized within wave for each respondent. We will use ten different headlines, balanced between five TRUE and five FALSE, and also balanced on the political orientation of the headlines.

- **Misinformation Literacy:** For this survey question from the recent Reuters Trust in Media Report (57). We use the following prompt: *How likely are you to do any of the following in the future?* with options ranging from Not at all likely to very likely. Items are described below:

  - Not sharing a news story when I am unsure about its accuracy
  - Checking a number of different sources to see whether a news story is reported in the same way
  - Relying on sources of news that are considered more reputable
  - Stopping to use certain news sources when I am unsure about the accuracy of their reporting
  - Discussing a news story with a person I trust because I am unsure about its accuracy
  - Stopping to pay attention to news shared by someone because I am unsure whether I trust that person

- Feeling Thermometer: Measure the respondent's feeling on a thermometer scale of 0 (negative) to 100 (positive). The value of the thermometer is assigned to the respective variable. We ask participants to rate the following options?
  - Democratic Party
  - Republican Party
  - News Media Organizations
  - Donald Trump
  - Joe Biden
- **Media Trust** Measures the respondent's trust in media using the summed value assigned to the respondent's answer to the next three questions. Questions wording and choices are presented below:
  - In general, how much trust and confidence do you have in the mass media – such as newspapers, TV and radio – when it comes to reporting the news fully, accurately and fairly
    - Not at all (0)
    - Not too much (0.5)
    - Some (0.75)
    - A lot (1)
  - In presenting the news dealing with political and social issues, do you think that news organizations deal fairly with all sides, or do they tend to favor one side? Which position is closer to your opinion?
    - Deal fairly with all sides (1)

    - Tend to favor one side (0)

- Based on what you know, how often do you believe the nation's major news organizations fabricate news stories?
    - All the time (0)
    - Most of the time (0.25)
    - About half of the time (0.5)
    - Once in a while (0.75)
    - Never (1)
    - **Political Cynicism:** Measures the respondent's levels of cynicism about democracy and political elites. Questions wording and choices are presented below:
    - Do you think that quite a few of the people running the government are corrupt, not very many are, or do you think hardly any of them are corrupt?
        - Quite a few (3)
        - Not very many (2)
        - Hardly any (1)
        - Don't know (NA)
    - Please indicate the extent to which you agree with the following statement: "Even though we live in a democracy a few people will always run things anyway."
        - Strongly disagree
        - Somewhat disagree
        - Neither agree nor disagree
        - Somewhat agree

        - Strongly agree

    - Social media news skepticism (1=strongly disagree to 7=strongly agree; 58, 59)
        - You cannot trust the news stories people share on social media
        - Too often credible news information is mixed up with misinformation on social media
        - You should not rely on social media for news information.

- Perceived ability to detect false rumors (scale 1-100; 53)
    - How do you think you compare to other Americans in your general ability to recognize news that is made up? Please respond using the scale below, where 1 means you're at the very bottom (worse than 99% of people) and 100 means you're at the very top (better than 99% of people)
    - How do you think you compare to other Americans in how well you performed in this study at recognizing news that is made up? Please respond using the scale below, where 1 means you're at the very bottom (worse than 99% of people) and 100 means you're at the very top (better than 99% of people).

Appendix B: Treatment Timeline

Our experiment consists of using the list feature from Twitter to create a pre-curated timeline formed by eight active fact-checking organizations with Twitter accounts on users' home feeds. In the pilot, we named the timeline as Fact-Checking Timeline. Figure 7 presents visually the intervention as we propose to deploy in the stage 2 report highlighting the new timeline in this visual representation.

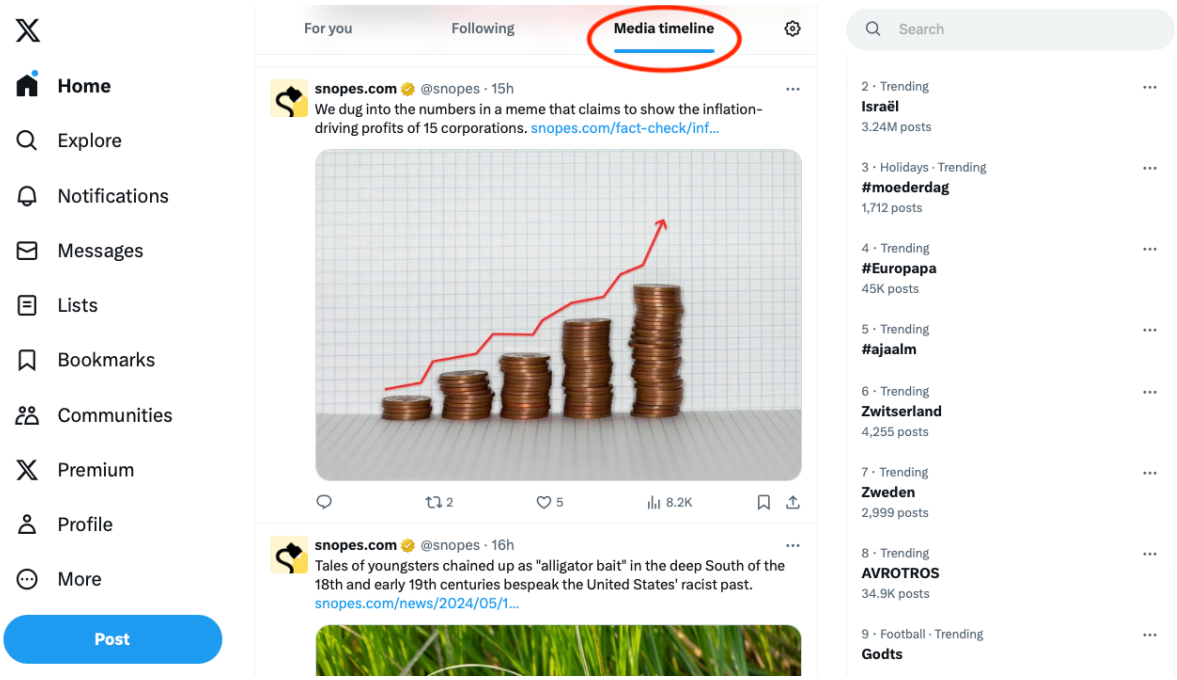


Figure 7 Example of the Timeline with Fact-Checking Organizations

Appendix C: Headlines Task

To study the effect of sustained exposure to fact-checking information, we will use a task using a set of true news stories and false rumors that circulated online during the period of the experiment. Respondents repeat this task ten times. For each headline, we ask participants:

“Please indicate whether you believe the following statements are accurate or not" to which respondents could choose from five responses: (1) Not at all accurate, (2) Not very accurate, (3) Somewhat accurate, (4) Very accurate, (5) Don’t know. True headlines will be collected from mainstream news media outlets, and false headlines will selected through salient misinformation corrected by multiple fact-checking organizations included in our intervention. All headlines will be selected from news published/corrected during the duration of the treatment.

Table 3 presents the full wording for the ten headlines used in the pilot study. We will select a new set of headlines for the full study using the same procedure described in this section.

Table 3 Pilot Study: True and false headlines selected to measure truth discernment.

Veracity	Headlines in English
False Item 1	The judge recently assigned to Donald Trump's classified documents case, was photographed at a Trump for President rally.
False Item 2	The blue, pink, and white colors in the progress pride flag represent pedophiles
False Item 3	Lindsay Graham recently told Ukrainian President Volodymyr Zelensky "The Russians are dying," and that is "the best money" the US has ever spent.
False Item 4	Many members of Joe Biden's cabinet have not sworn an allegiance to the Constitution
False Item 5	Arizona is banning the use of electronic voting machines in the state ahead of the 2024 presidential election
True Item 1	Ukrainian Military Drones Hit Residential Buildings in Moscow in May 2023
True Item 2	Last month Fox News chyron about Joe Biden read, "Wannabe dictator speaks at the White House after having his political rival arrested".
True Item 3	Trump did not sign any major immigration laws, while he was president
True Item 4	As part of a drug harm-reduction initiative, New York City launched a vending machine program offering free "safer smoking kits"
True Item 5	When Donald Trump was president, he signed a law changing the penalty for mishandling classified docs to 5 years in prison made it a felony

Appendix D: Minimum Detectable Effect Power Analyses

Figure 8 Minimum Detectable Effects for different combinations of observations and statistical power of 95% for Hypothesis 1: Truth Discernment (unadjusted ITT Model). Red line denotes effect size that we measured in our pilot study using the unadjusted ITT model.

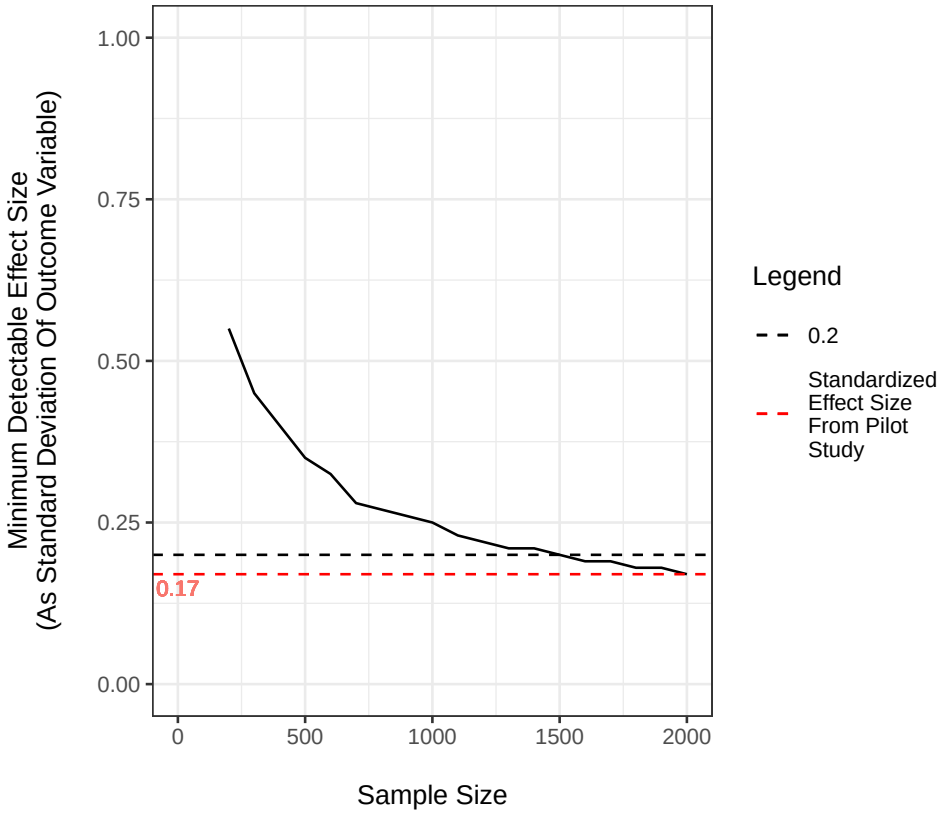


Figure 9 Minimum Detectable Effects for different combinations of observations and statistical power of 95% for Hypothesis 2: Misinformation Literacy (unadjusted ITT Model). Red line denotes effect size that we measured in our pilot study using the unadjusted ITT model.

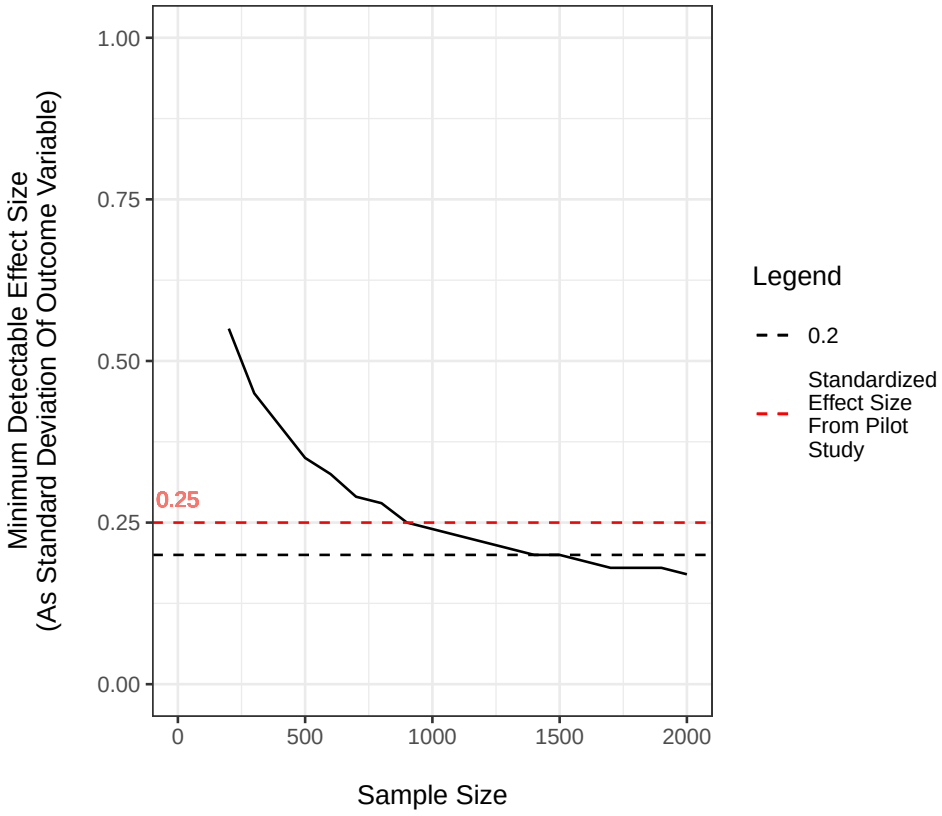


Figure 10 Minimum Detectable Effects for different combinations of observations and statistical power of 80% for Hypothesis 3: Conspiratorial Predispositions (unadjusted ITT Model). Red line denotes effect size that we measured in our pilot study using the unadjusted ITT model.

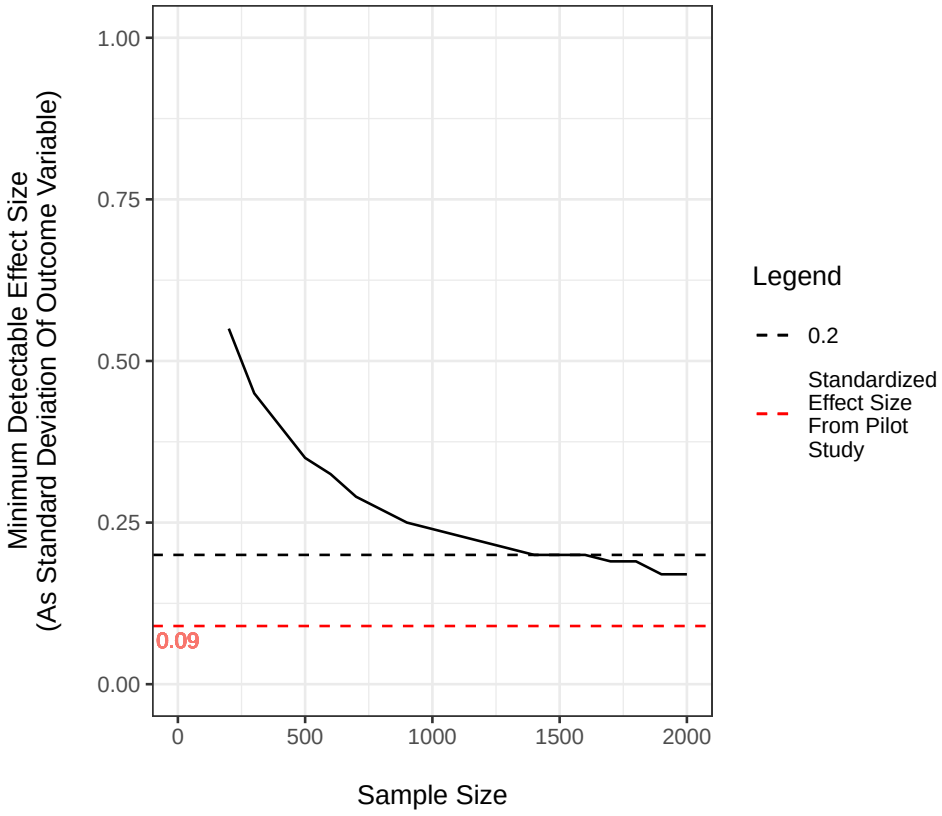


Figure 11 Minimum Detectable Effects for different combinations of observations and statistical power of 95% for Hypothesis 4: Affective Polarization (Party) (unadjusted ITT Model). Red line denotes the effect size that we measured in our pilot study using the unadjusted ITT model.

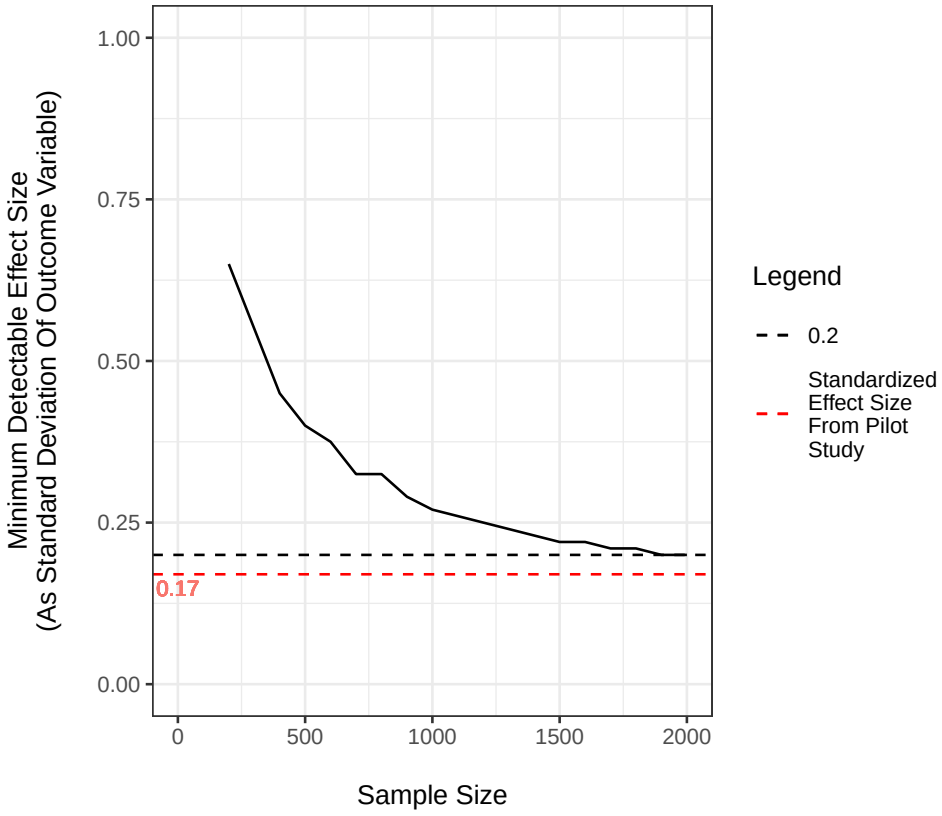


Figure 12 Minimum Detectable Effects for different combinations of observations and statistical power of 95% for Hypothesis 4: Media Trust (unadjusted ITT Model). Red line denotes effect size that we measured in our pilot study using the unadjusted ITT model.

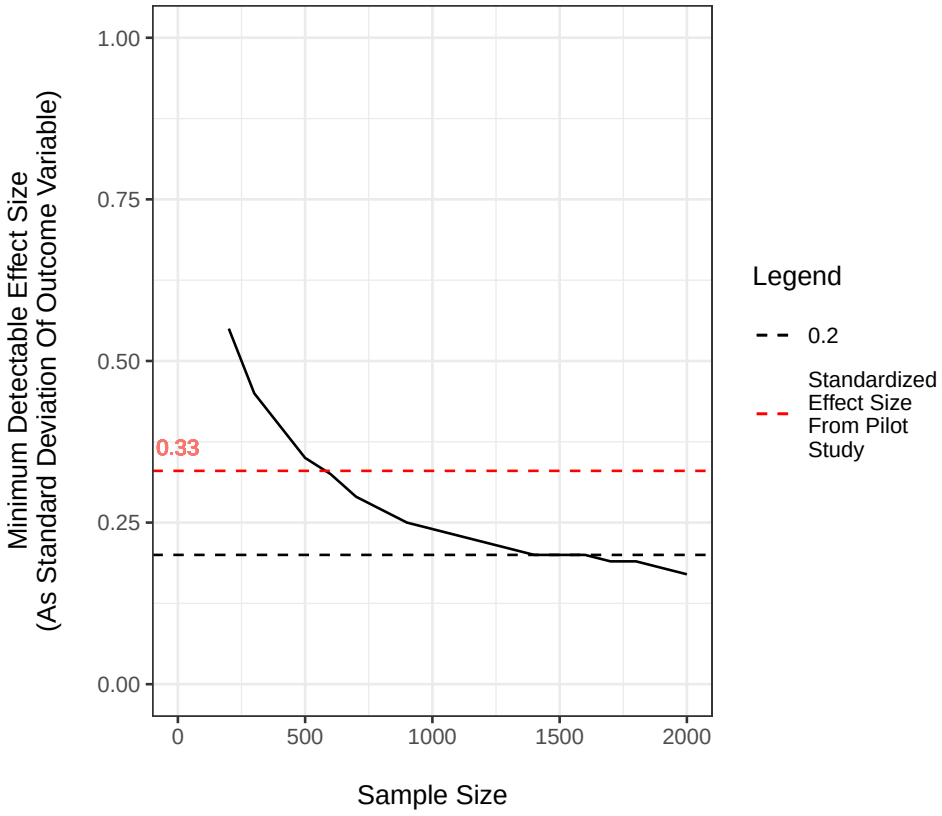


Figure 13 Minimum Detectable Effects for different combinations of observations and statistical power of 95% for Hypothesis 5: Political Cynicism (unadjusted ITT Model). Red line denotes effect size that we measured in our pilot study using the unadjusted ITT model.

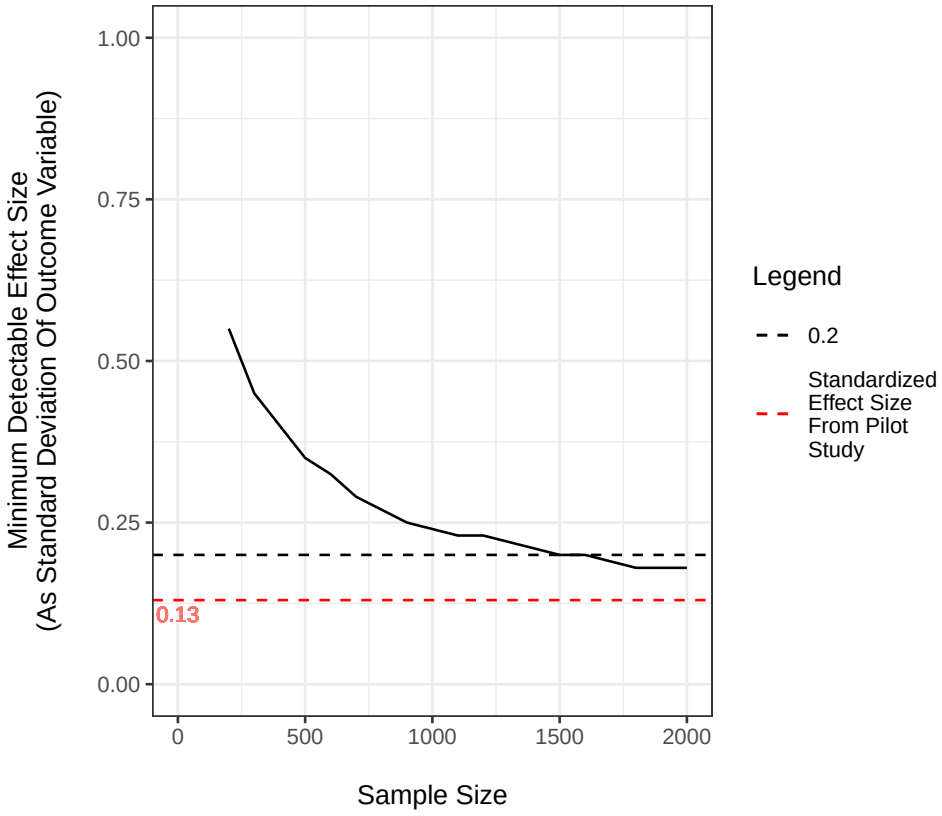
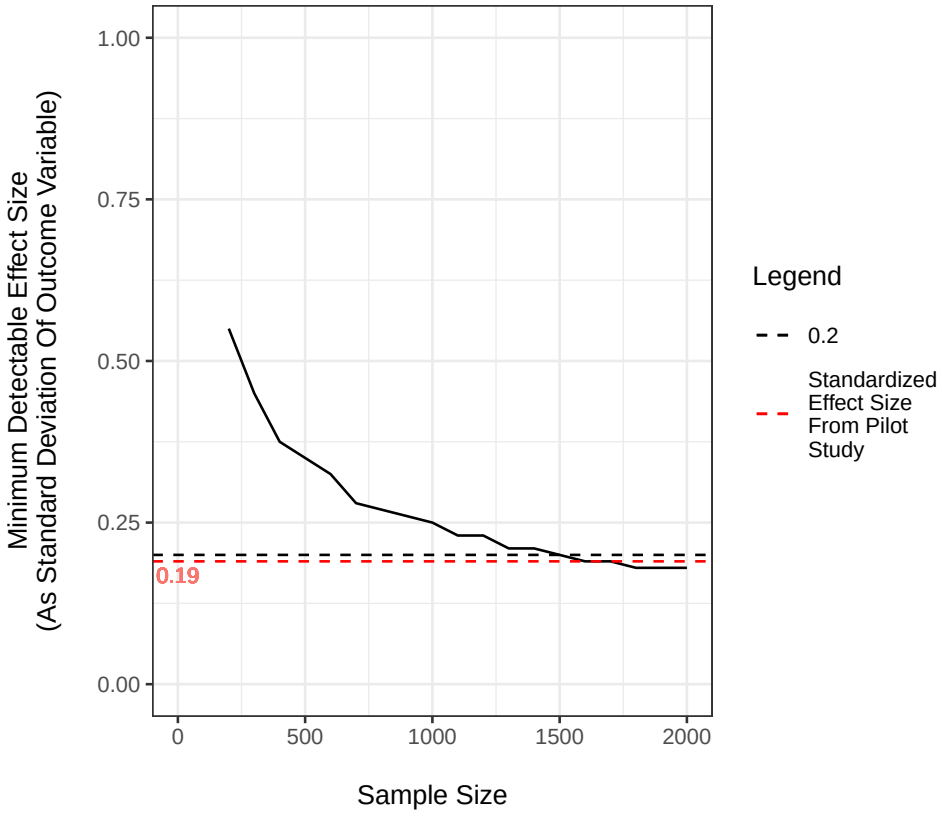


Figure 14 Minimum Detectable Effects for different combinations of observations and statistical power of 95% for Hypothesis 6: Social Media Cynicism (unadjusted ITT Model). Red line denotes effect size that we measured in our pilot study using the unadjusted ITT model.



Appendix E: Survey Materials

Consent Form

Consent Form for IRB-FY2023-6870 Consent Form for IRB-FY2023-6870

You are invited to participate in a survey on national, and public affairs conducted by researchers at New York University.

This study will be conducted by Joshua Tucker, FAS - Politics, Arts & Science, New York University.

This survey aims to obtain information about the consumption of news and fact-checking information on X/Twitter and opinions on current events. During the study, you will be asked several questions about current events. You will also be asked to research and evaluate the true and false information that will be presented to you. Some of the claims shown to you may contain false or misleading information. Your participation is voluntary. You are free not to answer any questions or to withdraw from the study at any time. No known risks are associated with your participation in this research beyond those of everyday life.

Today, you will be invited to respond to a quick five minutes questionnaire. If you pass some prerequisites, we will email you an invitation to join our main study in a few weeks. Participants who utilize a VPN tool to hide the location of where they are taking the survey will be screened out of the study.

In this study, you will be randomly assigned to either of the options below:

Option 1:

You will be invited to complete a new 8-10-minute questionnaire in a few weeks.

Then, as part of our study, we will create a new timeline on your Twitter with a list of media organizations at the top of your Twitter newsfeed. Content produced by these accounts will also appear on the “for you” and “following” feeds. The media organizations added in this timeline

work by verifying the accuracy of claims made in public discourse, whether by politicians, media outlets, or other sources and publishing their assessment.

This condition will last for one month. During this time, you will be asked to take screenshots or links of your new media timeline every other day and upload that screenshot with an explanation of the tweet you viewed (2 minutes of effort per screenshot).

After one month, you will be invited to complete another 10-15-minute follow-up survey. You will be compensated with gift cards for each stage of the research. Your compensation will be the following:

- Initial survey: $ 2
- Screenshots/links:

  At least 10 screenshots/links: $ 18

  Between 5 and 10 screenshots/links: $ 13

  Less than 5 screenshots: no additional compensation
- Final survey: $ 5

In summary, if you complete all tasks requested for this group, you can be compensated up to $25.00.

Option 2:

You will be invited to complete a new 8-10-minute questionnaire in a few weeks.

No changes to your Twitter account will be made if you are assigned to this option.

You will be asked to take screenshots or links of your Twitter newsfeed every other day and upload that screenshot with an explanation of the tweet you selected (2 minutes of effort per screenshot).

After one month, you will be invited to complete another 10-15-minute follow-up survey.

You will be compensated with gift cards for each stage of the research. Your compensation will be the following:

- Initial survey: $ 2
- Screenshots-links:

  At least 10 screenshots/links: $ 8

  Between 5 and 10 screenshots/links: $ 3

  Less than 5 screenshots: no additional compensation Final survey: $5

In summary, if you complete all tasks requested for this group, you can be compensated up to $15.00.

Your compensation will be delivered at the closing of each survey. The first payment will be due at the completion of the first survey. The second payment will be due after the screenshots and the second survey are completed.

You may choose not to answer any or all questions. Furthermore, you can opt out of the study at any time without penalty. This study's duration is about four weeks. You will be helping to further the scientific understanding of attitude formation and the flow of political information. Identifying information about you will never be used in any presentation or publication written about this project.

The confidentiality of your research records will be strictly maintained. Once anonymized, survey data will be stored on secure NYU servers.

Participation in this study is voluntary. You may refuse to participate or withdraw at any time without penalty. For interviews, questionnaires, or surveys, you have the right to skip or not answer any questions you prefer not to answer.

We will collect your public Twitter data that could contain identifying information, but it will not be linked to your survey data. The public Twitter data we collect will be used to generate

a simple set of non-identifiable measures, which will only then be connected to the survey data. All identifiable data will be deleted at the end of the study. Therefore, the identity of any human subjects cannot readily be ascertained through identifiers linked to the subjects. Information not containing identifiers may be used in future research, shared with other researchers, or placed in a data repository without your additional consent.

To revoke your consent to any portion of this study, email nyutwittersurveys@gmail.com with your request. You have the option to reach out to us in the email provided if you decide you no longer wish to have the intervention on the Twitter account OR if you no longer would like researchers to collect your Twitter information.

If there is anything about the study or your participation that is unclear or that you do not understand, if you have questions or wish to report a research-related problem, you may contact Joshua Tucker at (212) 998-7598, 19 W. 4th Street, New York, NY, 10012.

For questions about your rights as a research participant, you may contact the University Committee on Activities Involving Human Subjects (UCAIHS), New York University, 665 Broadway, Suite 804, New York, New York, 10012, at ask.humansubjects@nyu.edu or (212) 998-4808. Please reference study (IRB-FY2023-6870) when contacting the IRB (UCAIHS).

You have received a copy of this consent document to keep.

Please select the option below to indicate your consent to participate in this study:

The purpose and nature of this research have been sufficiently explained, and I agree to participate in this study. I understand that I am free to withdraw at any time without incurring any penalty. I understand that I may print this page for my records.

I decline to participate in this research. I understand that this survey will close.

Recruitment Survey

In this section, we present the full questionnaire for the recruitment survey used in the pilot study. Compared to this questionnaire, we expect to bring some of the questions from the pre-treatment survey already to the recruitment phase. Otherwise, we do not anticipate any substantive modifications in the questionnaire for the full study.

Questionnaire

[q-age]: What is your age?

- Under 18
- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65+
- Prefer not to answer

[q-gender]: Which gender do you self-identify with?

- Male
- Female
- Other
- Prefer not to answer

[q-race] What is your race? (multiple choice)

- White
- Black or African American
- Native American or Alaska Native
- Asian
- Native Hawaiian or Pacific Islander
- Other Racial Group (open-ended)

[q-education]: What is the highest education level you obtained?

- Less than high school

- High school graduate
- Some college
- 2 year degree
- 4 year degree
- Professional degree
- Doctorate

[q-zipcode]: What is the zip code of the area you live in? - [open-ended]

[q-email]: What is your email address? Remember, we will use your e-mail address to send you the study instructions and rewards information. - [open-ended]

[q-email2]: Please confirm your email address by typing it again below - [open-ended]

[q-politics]: How closely do you follow politics on TV, radio, newspapers, or the Internet?

- Very closely
- Fairly closely
- Not very closely
- Not at all

[q-social-media]: In which social media applications do you have an active account? By active account, we mean you open this application at least once every day. [Select all that apply]

- Facebook
- Twitter
- TikTok
- Instagram
- Telegram
- WhatsApp
- Other

[q-social-media-usage, include only items select from q-social-media]: Among the social media applications you have an active account, how frequently do you use them? [Scale: 1 .At least 10 times a day, 2. Several times a day, 3. About once a day 4 3 to 6 days a week, 5. 1 to 2 days a week, 6. Every few weeks, 7. Don't Know] [Options: Social media application selected on question 6.]

[q-twitter]: If you have a Twitter account, how much time, on average, do you spend on Twitter per day? Use your last week as a reference.

- I don't have an account on Twitter
- Less than 10 minutes
- Between 10 and 30 minutes
- Between 30 minutes and 1 hour
- Between 1 and 2 hours
- Between 2 and 4 hours
- More than 4 hours
- More than 6 hours
- I do not check my Twitter every day

[q-twitter-device]: What device do you use the most to access Twitter?

- Personal computer
- Mobile phone using the Twitter app
- Mobile phone using the web app
- Mobile phone shared with someone else

[q-twitter-purposes]: How frequently do you use Twitter for the following tasks? Please respond using the following scale [1 .At least 10 times a day, 2. Several times a day, 3. About once a day 4 3 to 6 days a week, 5. 1 to 2 days a week, 6. Every few weeks, 7. Don't Know]

- To keep up with family and friends online
- To stay up-to-date with news.
- To stay up-to-date with sports.
- To follow and interact with politicians
- To follow and interact with celebrities
- To share my opinion about politics
- To share my opinion about what is going around me

[q-follow] Approximately how many people do you follow on Twitter? [open-ended]

[q-followers] Approximately how many followers do you have on Twitter? [open-ended]

[q-study]: Now, at the end of this survey, we would like to invite you to participate in a longer study

Participate in our study!

WE OFFER: Up to $25 in gift cards if you fill out every survey.

YOUR PART: This study will last one month. To participate, you must fill out two 15-minute surveys ($2 for the first and $5 for the second survey a month later). In addition, you have the opportunity to earn an extra $18 for sending screenshots of your Twitter feed over the month. You can complete up to 15 screenshots over the course of the month to earn $25 in total.

ABOUT US: We are a team of international researchers who study social media. Your participation in this study will help us study the effects of Twitter on political opinions in the United States.

For this study, we ask that you engage with Twitter and fill out short surveys about your experience. As part of the study, you will be asked to perform a few different activities on Twitter. For example, you might be asked to follow a few different media accounts, add a second timeline to your Twitter mobile, and/or send us a few screenshots about your timeline. To facilitate your participation in the study, we can make all these modifications directly to your Twitter account using the official Twitter API. To participate in this study, you will be asked to authorize our access to your Twitter API.

In a few weeks, we will re-contact you by email to explain in detail the rules and tasks of the study.

By answering below "Yes, count me in!" we will take you to an app developed by us to authorize our access to your Twitter API. Thank you.

- Yes, count me in!
- No, not interested.

[q-end]: Thank you for providing us the Twitter authorization. We will recontact you in a few weeks with more information about the one-month study.

Move to the next page to finalize the study.

Pre-Treatment Survey

In this section, we present the full questionnaire for the pre-treatment survey used in the pilot study. This survey contains both the treatment revelation and measurement of pre-treatment outcomes. For the pilot, we first presented participants with their treatment assignment, and then asked pre-treatment questions. Because such ordering might cause post-treatment bias in the first measurement of the outcomes, in the full study, the treatment revelation will be performed at the end of the study.

[Treatment] Instructions for Treatment

Thank you for continuing your participation in our study! This study is conducted by researchers at the Center for Social Media and Politics at New York University.

You were assigned to Media Timeline group. By being part of this study, a new timeline of fact-checking organizations will be added to your Twitter feed.

This new timeline is composed of other Twitter accounts from reputable fact-checking organizations in the United States. These organizations verify the factual accuracy of questionable reporting and statements made by politicians, journalists, government officials, among others. After you finish this questionnaire, it will take 24 hours for the timeline to appear on your Twitter feed.

Your Twitter feed will be similar to the image below.

If you complete all the tasks of this study, you can earn up to $25 at the end of the research.

[new page]

Remember: You must regularly visit the Media Twitter timeline. To be sure you are paying attention to this new timeline we added to your Twitter feed, we will ask you every day to send us examples of the tweets that appear in your new timeline.

For this reason, over the course of four weeks, we will ask you to send us screenshots/links of tweets that appeared in the Media Timeline.

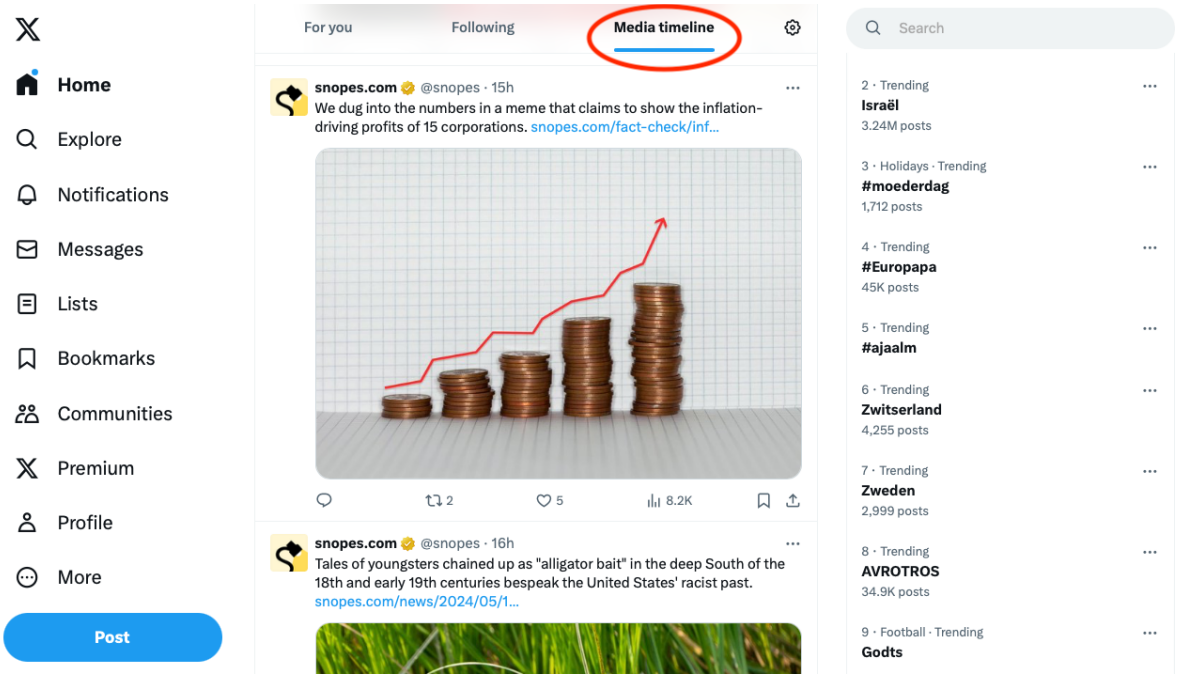


Figure 15 Example of Media Timeline

You can send us only one screenshot/link per day.

Tomorrow, you will get the first email with information about these tasks. Your final compensation for the study depends on how many of the following tasks you are able to complete:

- Finish survey today: $2
- Days that you send us screenshots/links:

  At least 10 days: $18 or

  Between 5 and 10 days: $13 or

  Less than 5 days: no additional compensation
- Finish the final survey: $5

As an active participant, you can earn up to $25!

If you would like to continue in the study, please respond yes to the question below. If you are not interested, say no, and you will receive your $2 compensation for your participation so

far.

- Yes, count me in!
- No, not interested.

[Control:] Instructions for Control

Thank you for continuing your participation in our study! This study is conducted by researchers at the Center for Social Media and Politics at New York University.

You were assigned to the Twitter Timeline group. In this group, as part of this study, you will tell us about posts you see on your Twitter timeline.

Remember: You need to send us Tweets from your timeline regularly To be sure you are paying attention to your Twitter account, we will ask you to send us examples of the tweets that appear in the timeline. For this reason, over the course of four weeks, we will ask you to send us screenshots/links of tweets that appeared in your Twitter Feed. You can send us only one screenshot/link per day.

Tomorrow, you will get the first email with information about these tasks. Your final compensation for the study depends on how many of the following tasks you are able to complete:

- Finish survey today: $2
- Days you send us screenshots/links:

  At least 10 days: $8 or

  Between 5 and 10 days: $5 or

  Less than 5 days: no additional compensation
- Finish the final survey: $5

As an active participant, you can earn up to $15!

If you would like to continue in the study, please respond yes to the question below. If you are not interested, say no, and you will receive your $2 compensation for your participation so far.

- Yes, count me in!
- No, not interested.

Digital-literacy: Please indicate your agreement with the following statements on a scale of Strongly Disagree to Strongly Agree.

- I prefer to ask friends how to use any new technological gadget instead of trying to figure it out myself.
- I feel like information technology is a part of my daily life.
- Using information technology makes it easier to do my work.
- I often have trouble finding things that I've saved on my computer.

q-trust: How much, if at all, do you trust the information you get from [scale:Not at all, Not too much, Some, A Lot]

- Newspapers
- Social media
- News websites
- Radio news
- Local news organizations
- National news organizations

q-ideo-place-self: When it comes to politics, would you describe yourself as liberal, conservative, or neither liberal nor conservative?

- Very liberal
- Somewhat liberal
- Slightly liberal
- Moderate; middle of the road
- Slightly conservative
- Somewhat conservative
- Very conservative

party-id Generally speaking, do you usually think of yourself as a Republican, a Democrat, an independent, or what?

- Republican
- Democrat

- Independent
- Something Else

party-id-lean, only if answered Independent or Something Else Do you think of yourself as closer to the Republican Party or to the Democratic Party?

- Closer to the Republican Party
- Closer to the Democratic Party
- Neither

strong-democrat, only if answered "Democrat": Would you call yourself a strong Democrat or a not very strong Democrat?

- Strong Democrat
- Not very strong Democrat

strong-republican, only if answered "Republican" Would you call yourself a strong Republican or a not very strong Republican?

- Strong Republican
- Not very strong Republican

social-media-cynicism Please indicate your agreement with the following statements on a scale of -4 = Strongly Disagree to 4 = Strongly Agree.

- You cannot trust the news stories people share on social media
- Too often credible news information is mixed up with misinformation on social media
- You should not rely on social media for news information.

Polcyn1: Do you think that quite a few of the people running the government are corrupt, not very many are, or do you think hardly any of them are corrupt?

- Quite a few (3)
- Not very many (2)
- Hardly any (1)
- Don't know (NA)

Media-trust-1: Some people think that by criticizing leaders, news organizations keep political leaders from doing their job. Others think that such criticism is valuable because it stops

political leaders from doing things that are not in the public's interest. Which position is closer to your opinion?

- Keeps political leaders from doing their job
- Stops political leaders from doing things that shouldn't be done

Media-trust-2: In presenting the news dealing with political and social issues, do you think that news organizations deal fairly with all sides, or do they tend to favor one side?

- Deal fairly with all sides
- Tend to favor one side

Media-trust-3: Based on your knowledge, how often do you believe the nation's major news organizations fabricate news stories?

- All the time
- Most of the time
- About half the time
- Once in a while
- Never

fc-familiarity: Are you familiar with the fact-checking practice in journalism, which includes websites such as PolitiFact, Factcheck.org, and the Washington Post Fact Checker? (Fact-checking is a development in journalism that seeks to hold politicians accountable when they make false or misleading statements.)

- Yes
- No

fc-trust: How much trust and confidence do you have in the fact-checking practice in journalism?

- None
- Little
- A moderate amount
- A lot
- A great deal

fake-news-problem: How much of a problem do you think made-up news and information is in the country today?

- A very big problem
- A moderately big problem
- A small problem
- Not a problem at all

headlines-task: Below are statements people have been making about issues of national relevance. Please indicate whether you believe the following statements are accurate or not. [Options: Not at all accurate, Not very accurate, Somewhat accurate, Very accurate, Don't know]

[Headlines will be selected according to the method described in Appendix C]

- Not at all accurate
- Not very accurate
- Somewhat accurate
- Very accurate
- Don't know

Misinfo-literacy How likely are you to do any of the following when navigating social media and news websites? [Seven Point Scale: Not at all likely ... Very likely]

- Not sharing a news story when I am unsure about its accuracy
- Checking a number of different sources to see whether a news story is reported in the same way
- Relying on sources of news that are considered more reputable
- Stopping to use certain news sources when I am unsure about the accuracy of their reporting
- Discussing a news story with a person I trust because I am unsure about its accuracy
- Stopping to pay attention to news shared by someone because I am unsure whether I trust that person

Feeling-Towards-SM-Platforms-Thermometer How much do you trust the news you see on

the following social media platforms? [Thermomether from 0-100]

- Facebook
- Twitter
- Instagram
- YouTube

Compliance Survey

[Treatment Group]

Thank you for continuing to participate in our Twitter study!

Your participation is helping us to learn more about how people use Twitter in the United States.

To measure your level of participation in our study, you were previously informed that we will request evidence of your engagement throughout the four-week study. For this reason, you will have the option to send in a Twitter screenshot or link each day of the study. The more days you send responses, the more rewards you will earn.

In this survey, we'll teach you how to upload a Twitter screenshot or send us a link of a tweet. But first, here are some important reminders about this study.

Reminder 1:

By being part of this study, you have committed to:

- Allowing us to add a timeline of media organizations to your Twitter account - Visiting your new Media Twitter timeline regularly

Your commitment to this study is to participate for a month. You will receive email reminders with links to submit screenshots and, most importantly, to complete the final questionnaire. Your participation will aid our research and guarantee your gift card reward.

Reminder 2: Our commitment to you:

At the end of the four weeks, we will invite you to complete a 10-minute questionnaire. If you follow the study rules: submit Twitter screenshots or links through survey links you receive via email, and answer the final questionnaire, you will receive a gift card. The amount of the gift card depends on how many of the following tasks you are able to complete:

- Initial survey: $2 Screenshots/links:
- At least 10 screenshots/links: $18

- Between 5 and 10 screenshots/links: $13
- Less than 5 screenshots: no additional compensation
- Final survey: $5

As an active participant, you can earn up to $25 (by completing both surveys and submitting at least 10 Twitter screenshots or links). You will only be compensated for the amount of tasks you complete successfully.

Reminder 3: About our rules

To make sure you are aware of the study rules, here are a few reminders for you to read before sending us the Twitter screenshot or link. This will ensure you are eligible for rewards in this study. Read the survey instructions carefully. Make sure you are sending in Twitter screenshots or links. Do not remove the new timeline with the fact-checking organizations from your account until the study is complete. To receive credit, you must send a valid screenshot/link of a Tweet from the media timeline. We will monitor whether the screenshots and links sent are are Tweets. Thank you for your participation!

Proceed to read the instructions on how to send the Twitter screenshots or Twitter links.

QID8

You can decide whether you want to hand in a screenshot or a link to a tweet. Please select below. On the next page you will receive detailed instructions on how to submit the screenshot or link.

- Upload a screenshot (1)
- Insert a link (2)

If selected Upload a screenshot Please select one tweet from your new Twitter media timeline that you found most informative today. It can be any kind of tweet from your Timeline. You should make sure that the whole tweet is captured. To find out how to make a screenshot on your current device, you can check out the instructions given here https://www.take-a-screenshot.org/ (opens in new window). This is an example of a tweet:

Once you have taken the screenshot, you can upload it below.

QID26 Below you can preview the screenshot that you are submitting. If this is not a tweet, or not the tweet you meant to submit, you can go back and upload a different screenshot. If this is the correct tweet, please answer the question below.

QID27 Why did you select this tweet? Please describe your reason in 1-2 sentences.

[open-ended]

If selected insert a link Please select one tweet from your new Twitter media timeline that you found most informative today. It can be any kind of tweet. To find out how to get the link of a tweet, you can check out the instructions given here https://www.teetweets.com/blogs/digest/how-to-find-tweet-url (opens in new window). Once you have copied the link, you can insert it below.

QID18 Please insert the Tweet link

QID24 Below you can see the tweet belonging to the link that you inserted earlier. If this is not the tweet you meant to submit, you can now go back and change the link. If this is the correct tweet, please answer the question below.

QID10 Here is your Tweet

QID28 Why did you select this tweet? Please describe your reason in 1-2 sentences.

[open-ended]

[Control Group]

Thank you for continuing to participate in our Twitter study!

Your participation is helping us to learn more about how people use Twitter in the United States.

To measure your level of participation in our study, you were previously informed that we will request evidence of your engagement throughout the four-week study. For this reason, you will have the option to send in a Twitter screenshot or link each day of the study. The more days you send responses, the more rewards you will earn.

In this survey, we'll teach you how to upload a Twitter screenshot or send us a link of a tweet. But first, here are some important reminders about this study.

Reminder 1: You have committed to:

- Visit your Twitter profile regularly

Your commitment to this study is to participate for a month. You will receive email reminders with links to submit screenshots and, most importantly, to complete the final questionnaire. Your participation will aid our research and guarantee your gift card reward.

Reminder 2: Our commitment to you:

At the end of the four weeks, we will invite you to complete a 10-minute questionnaire. If you follow the study rules: submit Twitter screenshots or links through survey links you receive via email, and answer the final questionnaire, you will receive a gift card. The amount of the gift card depends on how many of the following tasks you are able to complete:

- Initial survey: $2 Screenshots/links:
- At least 10 screenshots/links: $8
- Between 5 and 10 screenshots/links: $3
- Less than 5 screenshots: no additional compensation
- Final survey: $5

You can thus in total earn up to $15 (if you complete both surveys and send in at least 10 screenshots/links). If you choose to only complete part of the study, you will only be compensated for the tasks you completed.

Reminder 3: About our rules To make sure you are aware of the study rules, here are a few reminders for you to read before sending us the Twitter screenshot or link. This will ensure you are eligible for rewards in this study. Read the survey instructions carefully. Make sure you are sending in Twitter screenshots or links. To receive credit, you must send a valid screenshot/link of a Tweet from your timeline. We will monitor whether the screenshots and links sent are are Tweets. Thank you for your participation!

Proceed to read the instructions on how to send the Twitter screenshots or Twitter links.

QID8

You can decide whether you want to hand in a screenshot or a link to a tweet. Please select below. On the next page you will receive detailed instructions on how to submit the screenshot or link.

- Upload a screenshot (1)
- Insert a link (2)

If selected screenshot Please select one tweet from your Twitter timeline that you found most informative today. It can be any kind of tweet. You should make sure that the whole Tweet is captured.

To find out how to make a screenshot on your current device, you can check out the instructions given here https://www.take-a-screenshot.org/ (opens in new window). This is an example of a Tweet:

Once you have taken the screenshot, you can upload it below.

QID9 Please upload the screenshot here:

QID32 Below you can preview the screenshot that you inserted earlier. If this is not the tweet you meant to submit, you can now go back and change the screenshot. If this is the correct tweet, please answer the question below.

QID34 Why did you select this tweet? Please describe your reason in 1-2 sentences.

[open-ended]

If selected Insert a Link Please select one tweet from your Twitter timeline that you found most informative today. It can be any kind of tweet. To find out how to get the link of a tweet, you can check out the instructions given here https://www.teetweets.com/blogs/digest/how-to-find-tweet-url (opens in new window). Once you have copied the link, you can insert it below.

QID20 Please insert the Tweet link

QID29 Below you can see the tweet belonging to the link that you inserted earlier. If this is not the tweet you meant to submit, you can now go back and change the link. If this is the correct tweet, please answer the question below.

QID30 Here is your Tweet

QID31 Why did you select this tweet? Please describe your reason in 1-2 sentences.

[open-ended]

Post-Treatment Survey

In this section, we present the full questionnaire for the post-treatment survey used in the pilot study.

cheating-misinfo: The next set of questions helps us learn what types of information are commonly known to the public. Please answer these questions on your own without asking anyone or looking up the answers. Many people don't know the answers to these questions, but we'd be grateful if you would please answer every question. It is important to us that you do NOT use outside sources like the Internet to search for the correct answer. Will you answer the following questions without help from outside sources?

- Yes
- No

Headlines-Tasks Please indicate whether you believe the following statements are accurate or not. [Choices: Not at all accurate, Not very accurate, Somewhat accurate, Very accurate, Don't know]

- The judge recently assigned to Donald Trump's classified documents case, was photographed at a Trump for President rally.
- The blue, pink, and white colors in the progress pride flag represent pedophiles
- Lindsay Graham recently told Ukrainian President Volodymyr Zelensky "The Russians are dying," and that is "the best money" the US has ever spent.

- Many members of Joe Biden's cabinet have not sworn an allegiance to the Constitution
- Arizona is banning the use of electronic voting machines in the state ahead of the 2024 presidential election
- When Donald Trump was president he signed a law changing the penalty for mishandling classified docs to 5 years in prison and made it a felony
- Ukrainian Military Drones Hit Residential Buildings in Moscow in May 2023
- Last month Fox News chyron about Joe Biden read, "Wannabe dictator speaks at the White House after having his political rival arrested".
- Trump did not sign any major immigration laws, while he was president
- As part of a drug harm-reduction initiative, New York City launched a vending machine program offering free "safer smoking kits"

trust How much, if at all, do you trust the information you get from ... [Choices: Not at all, Not too much, Some, A lot]

- Newspapers
- Social media
- News websites
- Radio news
- Local news organizations
- National news organizations

trust-mainstream-news: How much, if at all, do you trust the information you get from ... [Choices: Not at all, Not too much, Some, A lot]

- Fox News
- CNN
- Snopes
- Politifact
- Verifythis

social-media-usage: In the past month, how frequently did you use each of the social media platforms below? [Choices: 1 .At least 10 times a day, 2. Several times a day, 3. About once a day 4 3 to 6 days a week, 5. 1 to 2 days a week, 6. Every few weeks, 7. Don't Know, 8. Don't

have an account]

- Facebook
- Twitter

affective-polarization

We'd like to get your feelings toward political groups using a "feeling thermometer" A rating of 0 degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as possible. You will rate the group 50 degrees if you don't feel particularly positive or negative toward the group.

Please use the slider to enter the "degree" or number between 0 and 100 for your feeling about the following groups: (slider 0-100)

- Democratic Party
- Republican Party
- News Media Organizations
- Donald Trump
- Joe Biden

conspiracy-grid: Please indicate the extent to which you agree with the following statements: [Choices: Strongly disagree; Somewhat disagree ; Neither agree nor disagree; Somewhat agree; Strongly agree]

- Much of our lives are being controlled by plots hatched in secret places
- Even though we live in a democracy a few people will always run things anyway.
- The people who really 'run' the country are not known to the voter.
- Big events like wars, recessions, and the outcomes of elections are controlled by small groups of people who are working in secret against the rest of us

Misinfo-literacy How likely are you to do any of the following when navigating social media and news websites? [Choices: 1=Not at all likely; 7=very likely]

- Not sharing a news story when I am unsure about its accuracy
- Checking a number of different sources to see whether a news story is reported in the same way

- Relying on sources of news that are considered more reputable
- Stopping to use certain news sources when I am unsure about the accuracy of their reporting
- Discussing a news story with a person I trust because I am unsure about its accuracy
- Stopping to pay attention to news shared by someone because I am unsure whether I trust that person

social-media-cynicism

Please indicate your agreement with the following statements on a scale of Strongly Disagree to Strongly Agree.

- You cannot trust the news stories people share on social media
- Too often credible news information is mixed up with misinformation on social media
- You should not rely on social media for news information.

Pol-cyn-1: Do you think that quite a few of the people running the government are corrupt, not very many are, or do you think hardly any of them are corrupt?

- Quite a few (3)
- Not very many (2)
- Hardly any (1)
- Don't know (NA)

Media-trust-1: Some people think that by criticizing leaders, news organizations keep political leaders from doing their job. Others think that such criticism is valuable because it stops political leaders from doing things that are not in the public's interest. Which position is closer to your opinion?

- Keeps political leaders from doing their job
- Stops political leaders from doing things that shouldn't be done

Media-trust-2: In presenting the news dealing with political and social issues, do you think that news organizations deal fairly with all sides, or do they tend to favor one side?

- Deal fairly with all sides
- Tend to favor one side

Media-trust-3: Based on your knowledge, how often do you believe the nation's major news organizations fabricate news stories?

- All the time
- Most of the time
- About half the time
- Once in a while
- Never

attn-check-1 Paying attention and reading the instructions carefully is critical. President Obama is the youngest politician listed in the options below. If you are paying attention, please choose the youngest politician from the list.

- President Joe Biden
- President Trump
- President Obama
- Nancy Pelosi
- Mike Pence
- Don't Know

fc-trust: How much trust and confidence do you have in the fact-checking practice in journalism?

- None
- Little
- A moderate amount
- A lot
- A great deal

fake-news-problem: How much of a problem do you think made-up news and information is in the country today?

- A very big problem
- A moderately big problem
- A small problem
- Not a problem at all

fc-share How often do you share/post on your social media accounts (Twitter, Facebook,

for example) corrections published by Fact-Checking organizations?

- None
- Little
- A moderate amount
- A lot
- A great deal

feeling-platforms How much do you trust the news you see on the following social media platforms? (slides 0-100)

- Facebook
- Twitter
- YouTube

Politic-effic Please indicate the extent to which you agree with the following statements: (choices: Strongly disagree; Somewhat disagree ; Neither agree nor disagree; Somewhat agree ; Strongly agree)

- I feel confident that I can find the truth about political issues.
- If I wanted to, I could figure out the facts behind most political disputes
- There are objective facts behind most political disputes, and if you try hard enough you can find them.

Fc-ability-general How do you think you compare to other Americans in your general ability to recognize news that is made up? Please respond using the scale below, where 1 means you're at the very bottom (worse than 99% of people) and 100 means you're at the very top (better than 99

[slider 0-100]

fc-ability-specific How do you think you compare to other Americans in how well you performed in this study at recognizing news that is made up? Please respond using the scale below, where 1 means you're at the very bottom (worse than 99

[slider 0-100]

factchecking-feeling How did you find your overall experience with the Media Timeline in your Twitter feed?

- I disliked it a lot
- I disliked it a little
- I neither liked it nor disliked it
- I liked it a little
- I liked it a lot
- I don't know
- I didn't have the Media timeline in my Twitter feed.

[**Only for Treatment**]

fact-experience Below are statements about the media extension. Please say if you agree or disagree with each statement. (choices: Strongly disagree; Somewhat disagree ; Neither agree nor disagree; Somewhat agree ; Strongly agree)

- I barely noticed the Media Timeline when I was online.
- The Media Timeline was intrusive and bothered me while I was online.
- The Media Timeline corrections on Twitter are fair.
- The Media Timeline corrections about the news I visit frequently were more negative than I expected.

factchecking-use How likely are you to use the Media Timeline on your Twitter account after this study is finished?

- 1 - Not at all likely
- 7 - very likely

[All Participants]

party-id Generally speaking, do you usually think of yourself as a Republican, a Democrat, an independent, or what?

- Republican
- Democrat
- Independent
- Something Else

q-political-know1 For how many years is a United States Senator elected - that is, how many years are there in one full term of office for a U.S. Senator?

- Two years
- Four years
- Six years
- Eight years
- None of these
- Don't know

q-political-know2 How many times can an individual be elected President of the United States under current laws?

- Once
- Twice
- Four times
- Unlimited number of terms
- Don't know

q-lookup It is essential for the validity of this study that we know whether participants looked up any information online during the study. Did you make an effort to look up information during the study? Please be honest; you will still be paid and you will not be penalized in any way if you did.

- Yes
- No

compliance-treatment

In the last question of this questionnaire, we would like to offer you the opportunity to complete an additional task and win an additional two dollars on your final compensation.

We ask you to take a screenshot of your Twitter Timeline. This is not a screenshot of a Tweet. We want you to send us a screenshot of the top of your Twitter home, such as in the image below, where the 'for you' and 'following' and any additional timelines appear for you. To find out how to make a screenshot on your current device, you can check out the instructions

given here https://www.take-a-screenshot.org/ (opens in a new window)

Once you have taken the screenshot, you can upload it below.

If you prefer not to answer this question, you can skip to finish the questionnaire.

Appendix F: Pilot Results with Simulated Data

In this section, we present and discuss our pre-registered models using a simulate sample of 1,800 observations using bootstrap resampling methods and repeat this process 1,000 times to build credible intervals for predicted estimates of our specified models. This exercise intend to provide the reader with an approximation of the confidence intervals from the full manuscript using a larger sample size.

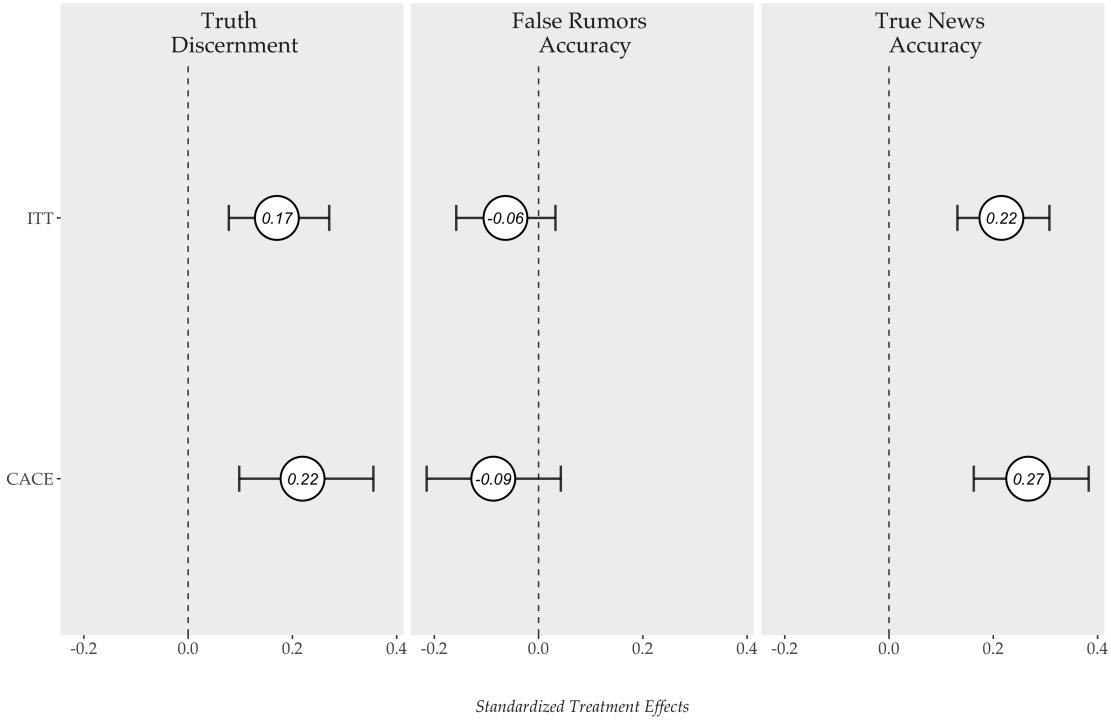


Figure 16 Simulated Treatment effects on the discernment of news veracity. Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% percentile bootstrapped confidence intervals

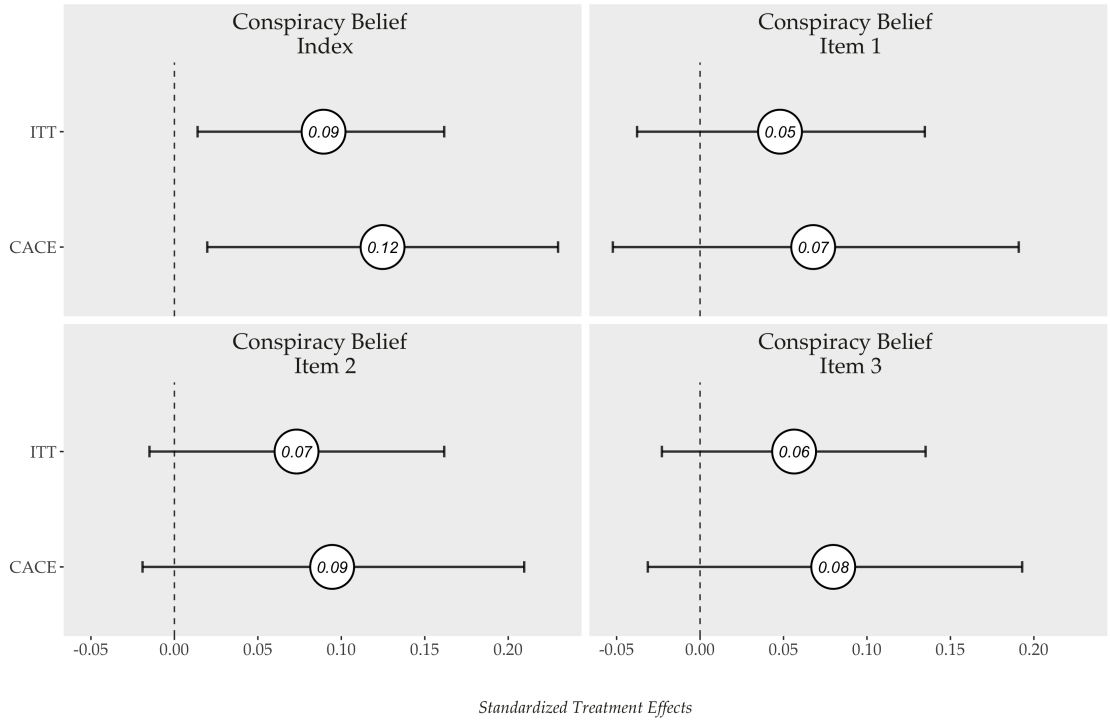


Figure 17 Simulated Treatment effects on conspiratorial predispositions. Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% percentile bootstrapped confidence intervals

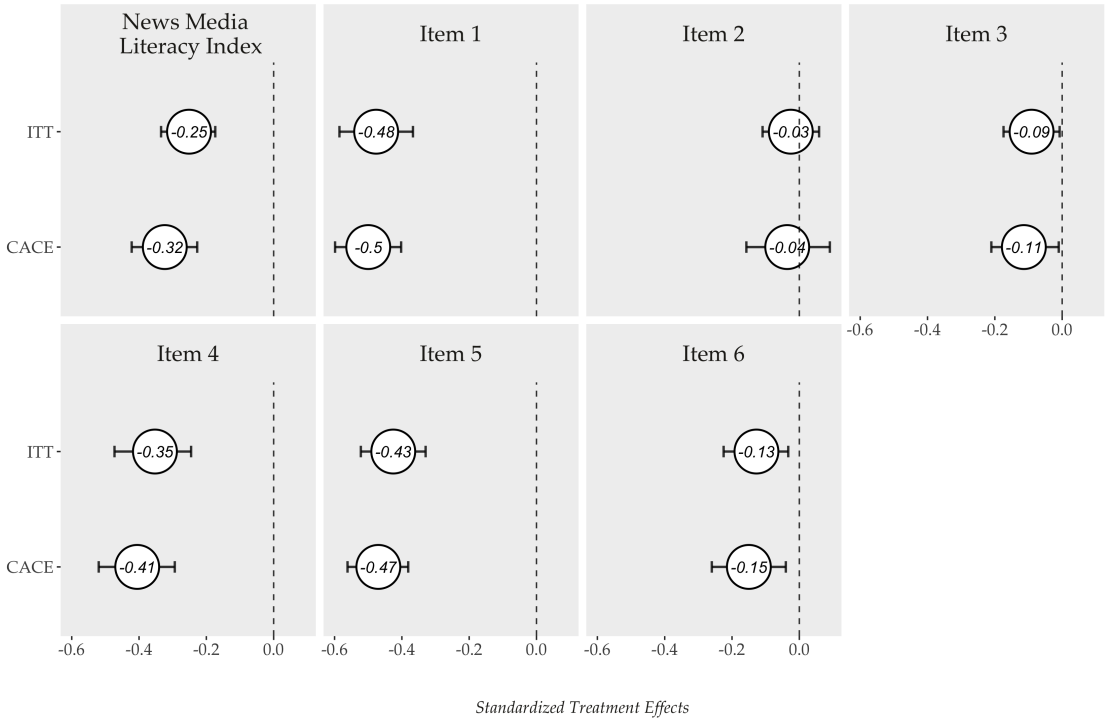


Figure 18 Simulated Treatment effects on Misinformation Literacy. Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% percentile bootstrapped confidence intervals

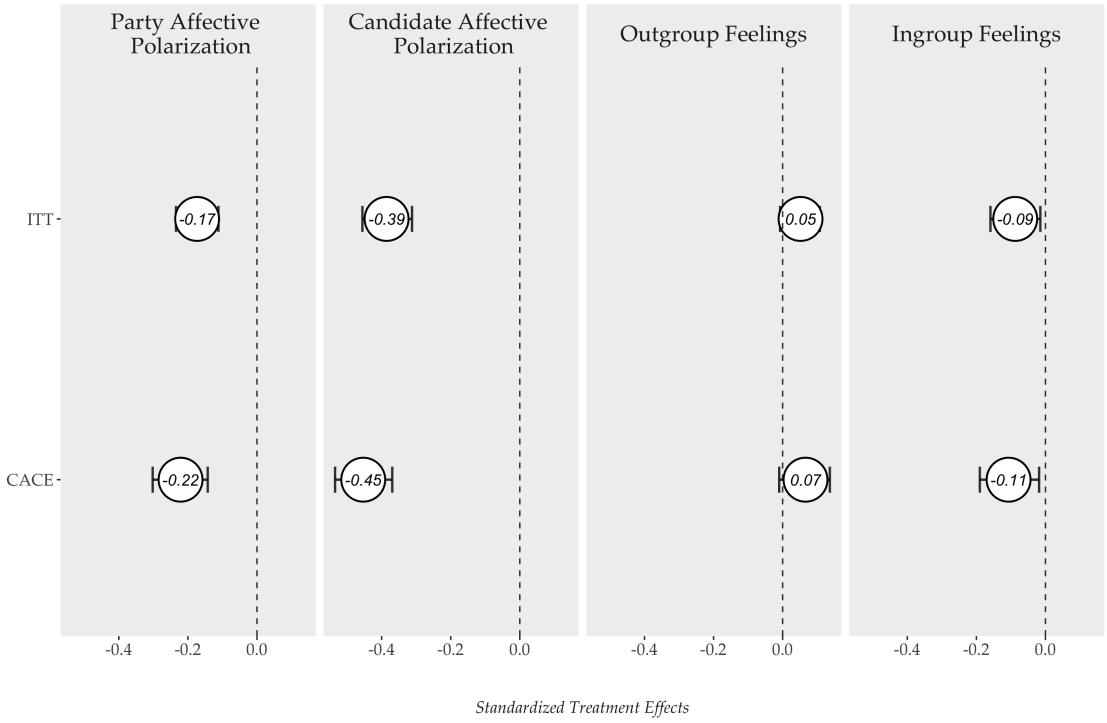


Figure 19 Simulated Treatment effects on Affective Polarization. Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% percentile bootstrapped confidence intervals

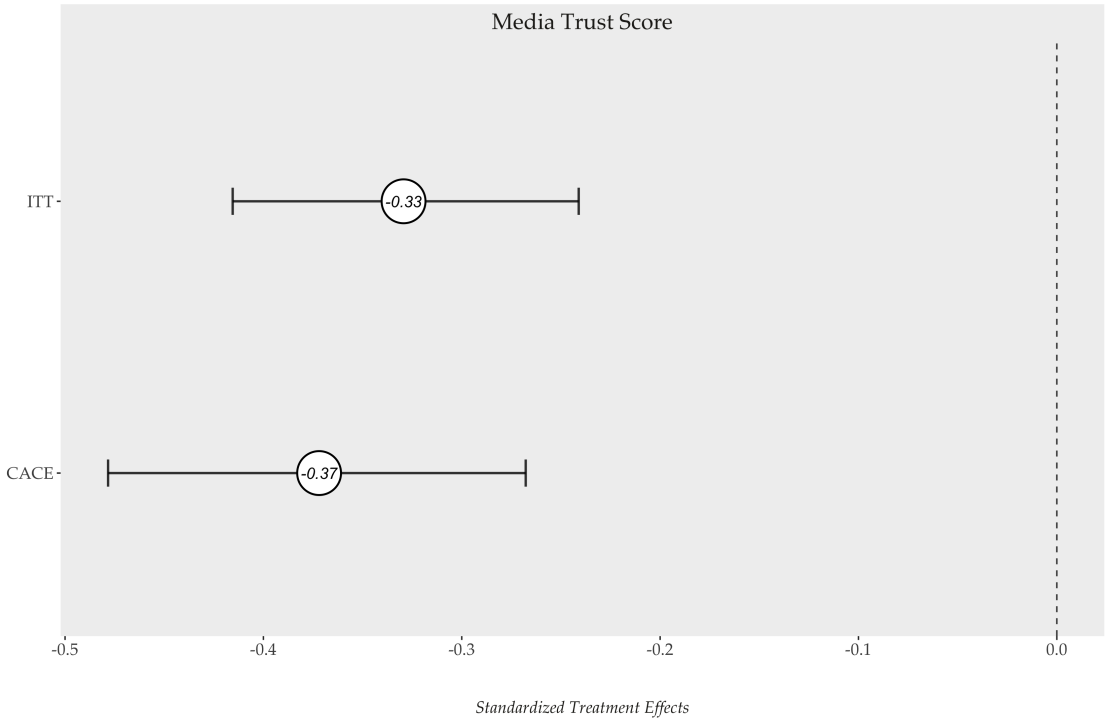


Figure 20 Simulated Treatment effects on Media Trust. Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% percentile bootstrapped confidence intervals

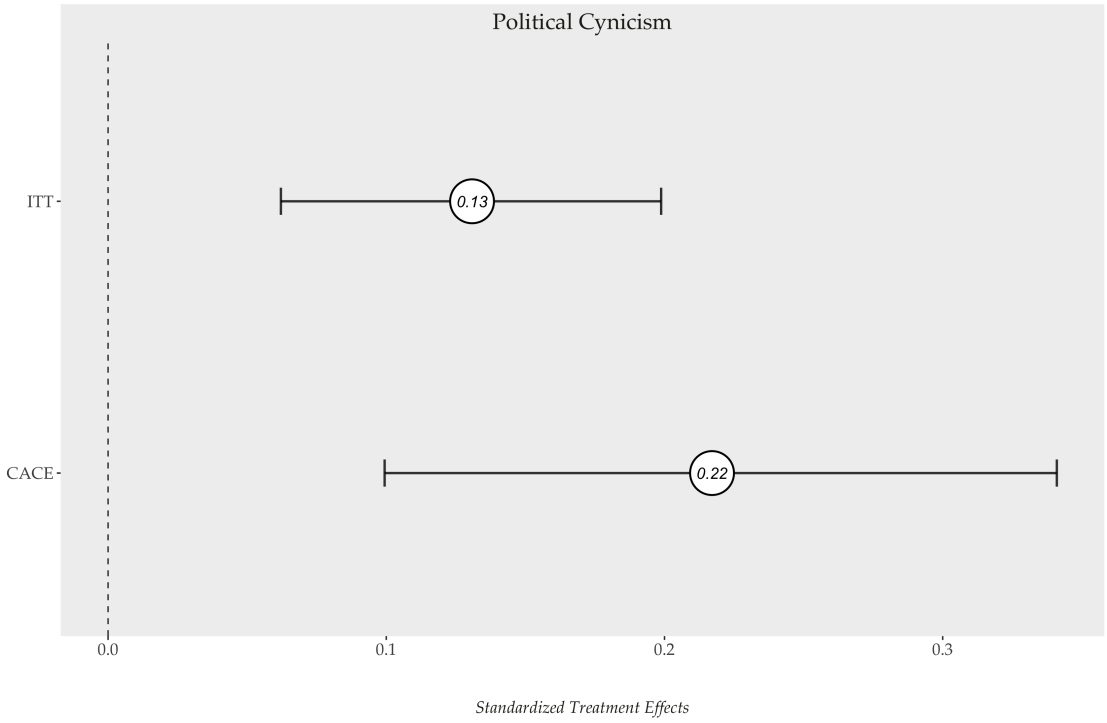


Figure 21 Simulated Treatment effects on Political Cynicism. Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% percentile bootstrapped confidence intervals

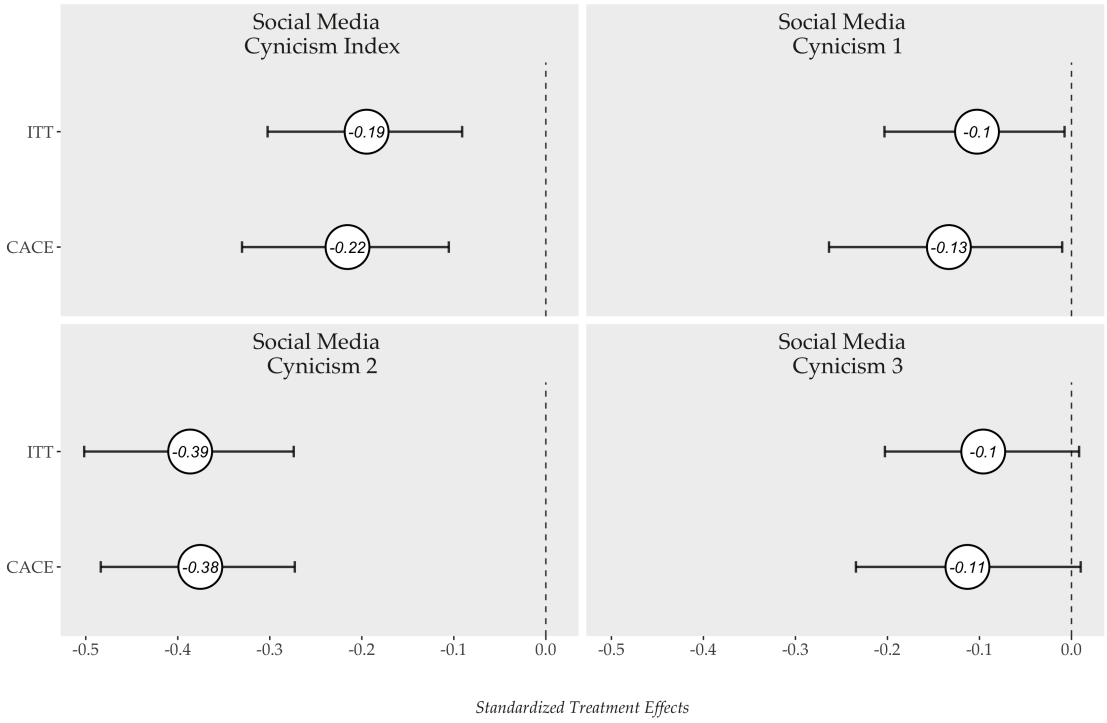


Figure 22 Simulated Treatment effects on Social Media Cynicism Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% percentile bootstrapped confidence intervals

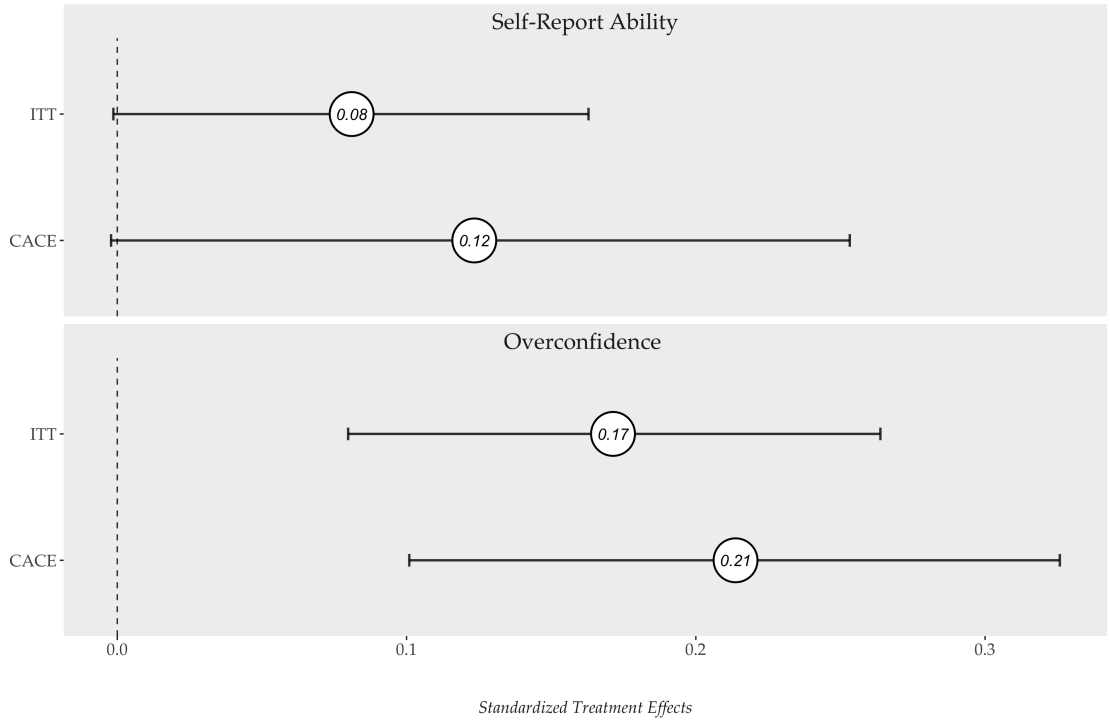


Figure 23 Simulated Treatment effects on Self-Reported and Overconfidence on Misinformation Judgements Point estimates are marginal effects relative to the outcome's standard deviation in the control group. The figure uses 95% percentile bootstrapped confidence intervals