Reducing Social Media Usage During Elections: Evidence from a Multi-Country WhatsApp Deactivation Experiment

Rajeshwari Majumdar[*,**]
Tiago Ventura[†,*]
Shelley Liu[‡]
Carolina Torreblanca[§]
Joshua A. Tucker[¶]

July 2025

Abstract

We deploy online field experiments in Brazil, India, and South Africa to examine how restricting the use of WhatsApp, the world's most widely used messaging app, affects information exposure, political attitudes, and individual well-being. We incentivize participants to either (1) stop consuming *multimedia* content on WhatsApp or (2) limit overall WhatsApp *usage* to 10 minutes per day for four weeks ahead of each country's elections. We find that our interventions significantly reduced participants' exposure to misinformation, online toxicity, and uncivil discussions about politics—but at the expense of keeping up with true political news. Using a wide range of measures, we detected no changes to political attitudes, but uncovered substantial gains to individual well-being as treated participants substituted WhatsApp usage for other activities. Results highlight the complex trade-offs associated with the effects of social media use on information consumption and its downstream effects.

[*]School of Media and Public Affairs, The George Washington University. r.majumdar@gwu.edu.
[†]McCourt School of Public Policy, Georgetown University. tv186@georgetown.edu.
[‡]Sanford School of Public Policy, Duke University. shelley.liu@duke.edu.
[§]PDRI-DevLab, University of Pennsylvania. catba@sas.upenn.edu.
[¶]Department of Politics and Center for Social Media and Politics, New York University. joshua.tucker@nyu.edu.

1 Introduction

Social media platforms have fundamentally reshaped political communication in the modern era, influencing not only how information is disseminated but also how individuals engage with politics. In their early years, these platforms created genuine enthusiasm among policymakers and experts about breaking informational barriers, facilitating connections, and reducing costs of mobilization (Tucker et al. 2017). However, more recently, these platforms have been frequently associated with possible negative externalities. This issue is particularly salient in politics, where social media usage has been widely conceived as affecting political attitudes and behavior—with deleterious consequences for democracy. A growing body of literature has therefore examined a wide range of societal impacts associated with social media usage, such as facilitating the spread of misinformation (Guess, Nyhan and Reifler 2018; Flaxman, Goel and Rao 2016), creating echo chambers (Guess et al. 2023; Sunstein 2018), increasing political polarization (Banks et al. 2021; Tokita, Guess and Tarnita 2021; Bail et al. 2018; Settle 2018), exposing users to uncivil interactions (Bor and Petersen 2022), and harming mental health (Tromholt 2016; Vanman, Baker and Tobin 2018; Hanley, Watt and Coventry 2019).

Scholarship on the effects of social media, however, faces two core limitations. First, most of our existing knowledge is still derived from studies of traditional feed-based social media platforms such as Facebook or X, which are particularly popular in the United States and other Western democracies. However, in many parts of the world, information exchange and daily life are intertwined largely through social messaging apps such as WeChat, Telegram, and WhatsApp (Newman et al. 2021; Batista Pereira et al. 2023). This distinction is important: in comparison to algorithmic feeds, content propagation on social messaging apps depends more heavily on users forwarding content in group chats and one-to-one conversations, which fundamentally shapes the forms of engagement people have with information that they see (Valenzuela, Bachmann and Bargsted 2021; Rossini, Stromer-Galley, Baptista and Veiga de Oliveira 2021). Furthermore, without a social graph structure, where creators/influencers cater to their follower base, the most viral information on WhatsApp circulates quasi-anonymously, lacking any metadata, and is often crafted for easy distribution across different groups and chats. Consequently, WhatsApp's informational environment is dominated not by text-based news articles and posts, but by eas-

ily shareable multimedia content (such as videos, images, and audio) that constitutes much of the misinformation on the platform (Burgos 2019; Avelar 2019; Resende, Melo, Reis, Vasconcelos, Almeida and Benevenuto 2019; de Freitas Melo et al. 2019; Machado et al. 2019; Garimella and Tyson 2018; Garimella and Eckles 2020).

Second, outside of Western contexts, there are often different sets of overlapping challenges to misinformation and information flow. These challenges may be capacity-driven, such as less robust fact-checking apparatuses or greater mobile data costs (Bowles, Larreguy and Liu 2020; Haque et al. 2020); they may also be political in nature, considering increased risks of violence, low political trust, or authoritarian politics and greater political control over media (Jones 2022; Persily and Tucker 2020; Cheeseman et al. 2020; Asimovic et al. 2021; Badrinathan, Chauchard and Siddiqui 2024). In response, scholars have increasingly sought to examine countries around the globe (see discussion in Blair et al. 2024); however, these efforts often focus more narrowly on interventions to tackle the spread of misinformation and the persistence of inaccurate beliefs rather than on understanding the impacts of popular social media platforms on information sharing more broadly (Badrinathan and Chauchard 2023; Bowles et al. 2025).

To address these gaps in knowledge, we implement a large-scale online field experiment around WhatsApp usage in three major Global South democracies—India, South Africa, and Brazil—ahead of the 2024 elections in these countries. As three of the largest democracies in Asia, Africa, and Latin America respectively, these three countries are substantively important cases to study. Moreover, across these countries, and in most of the Global South, WhatsApp plays a major role in everyday life as the largest social media and messaging platform. WhatsApp is the most used app for diverse tasks in these countries, from communicating with friends and conducting business to consuming news, including political and election-related content (Newman et al. 2024). In the political realm, WhatsApp groups have become powerful tools for propaganda and organization (Chauchard and Garimella 2022; Gil de Zúñiga, Ardèvol-Abreu and Casero-Ripollés 2021), news consumption (Valenzuela, Bachmann and Bargsted 2021; Newman et al. 2024), and spreading misinformation (Bowles et al. 2025; Ventura et al. 2025).

Our study investigates social media effects on (mis)information exposure and their downstream consequences on both political and non-political outcomes through two experimental interventions that incentivize participants to reduce WhatsApp usage in a critical political moment.

Our research design refines upon existing experimental research that has sought to identify the causal impact of social media on various outcomes through *deactivation*—where users are instructed to temporarily deactivate their social media accounts entirely, while a comparison group continues their usage as usual (Allcott et al. 2020, 2024; Arceneaux et al. 2023; Asimovic et al. 2021; Asimovic, Nagler and Tucker 2023).[1] Instead of full deactivation, however, we pursue a reduction incentive design—a partial deactivation as in Ventura et al. (2025)—for ethical and methodological reasons. We do so considering how WhatsApp is embedded in users' personal and professional lives in ways that feed-based social media platforms like Facebook and Twitter are not (Rossini, Stromer-Galley, Baptista and Veiga de Oliveira 2021; Gil de Zúñiga, Ardèvol-Abreu and Casero-Ripollés 2021). Ethically, fully removing people from WhatsApp—where they communicate with family and conduct business—could lead to harms that supersede the potential benefits granted by our design. Methodologically, a full deactivation would significantly shrink the subject pool to a subset of users who are distinct from the population of regular WhatsApp users, greatly hampering external validity.[2]

Our treatment arms approach partial deactivation in two ways. First, as information on WhatsApp, particularly misinformation and polarizing content (Garimella and Eckles 2020; Resende, Melo, Sousa, Messias, Vasconcelos, Almeida and Benevenuto 2019), is often delivered through multimedia content – i.e., images, videos, documents, and audio – we ask one group of treated participants to turn off the automatic download of multimedia content on WhatsApp (*Multimedia* arm). We further incentivize them to not access any media during a four-week period preceding election day. Second, we incentivize another group of participants to limit their WhatsApp usage to a maximum of 10 minutes per day during the same four weeks (*Time* arm). To ensure a comparable control group that is as willing and able as our treatment groups are to undertake these tasks, participants assigned to *Control* are asked to perform the same tasks as the treatment groups, but for a much shorter period of only three days. Taken together, our experiment aims to reduce participants' WhatsApp usage ahead of elections, which we theorize ought to reduce exposure to information—true and false—and negative political content circulating on

[1] The goal of a classic deactivation treatment is to approximate two counterfactual scenarios, one population where social media has been adopted and is widely used, and another population where social media does not exist.

[2] WhatsApp usage is high in all three countries, both as the dominant social media platform and for accessing news (Poushter 2024; Newman et al. 2021). Thus, a study that primarily captured users who can easily eliminate WhatsApp usage entirely would fail to represent the broader population in these contexts.

social media during elections.

We present four core findings that underscore the complex tradeoffs of social media usage. First, reducing WhatsApp usage consistently reduced participants' exposure to *both* misinformation rumors and true news circulating in the weeks before elections. Yet, despite these changes in information exposure, our treatments did not influence participants' accuracy judgments for either misinformation or true news. Second, reducing WhatsApp usage also decreased participants' exposure to online incivility and to low-quality or toxic political discussion. Third, despite reducing treated participants' exposure (mis)information and uncivil content, our treatments had no downstream effects on political attitudes, including affective (partisan) polarization, identity-based prejudice, issue polarization, and candidate favorability. Fourth, we also document positive non-political effects, such as increased in-person social interactions and a significant boost in subjective well-being. In sum, while our intervention failed to enhance political knowledge or reduce polarization, it did yield considerable benefits in terms of decreasing exposure to toxicity and improving well-being.

Our study contributes both substantively and methodologically to the existing literature on social media effects. First, given that our study was deployed one month prior to major elections in each country, it contributes to the existing literature on media, information access, and political attitudes during politically polarizing periods. High levels of polarization make individuals less likely to update their priors with new information, making directional motivations and identity or partisan cues more salient (Graham and Svolik 2020; Druckman, Peterson and Slothuus 2013). Considering the media environment, this dynamic leads voters to increase their overall demand for political content, as well as their reliance on aligned pro-attitudinal sources (Arceneaux and Johnson 2013; Prior 2007; Stroud 2011). In this context, while elections are a critical period in democratic politics, these are also times in which political attitudes become increasingly more resistant to change. Our null effects on a wide range of political attitudes, despite considerable reductions in exposure to (mis)information and uncivil political content, align closely with theories of media minimal effects under polarizing times, and resonate with findings from other recent social media field experiments deployed in election periods (Allcott et al. 2024; Wojcieszak et al. 2022; Nyhan et al. 2023; Guess et al. 2023; Ventura et al. 2025).

Second, our study expands the emerging literature on the causal effects of social media. To

date, most prior studies designed to identify these effects have done so through randomized deactivations of Facebook users (Hanley, Watt and Coventry 2019; Vanman, Baker and Tobin 2018; Hall et al. 2021; Allcott et al. 2020; Asimovic et al. 2021; Asimovic, Nagler and Tucker 2023) and with a strong focus on Western countries. Social media messaging apps have become, in recent years, the primary way through which online misinformation and inflammatory content circulates in major markets in the Global South. By assessing the effects of messaging platforms on the overall informational environment and downstream effects on political and non-political attitudes, our study contributes new insights to the field. In addition, methodologically, our study also provides a novel comparative understanding of the intricate dynamics of the effects of social media usage across major Global South countries.

Third, our research contributes to a growing literature focusing on social media messaging apps. A key methodological challenge in studying WhatsApp and similar platforms stems from data limitations inherent to an encrypted application. Consequently, much of the emerging academic literature on WhatsApp has been primarily descriptive, focusing either on strategies to collect data (Garimella and Tyson 2018; Garimella and Eckles 2020; Chauchard and Garimella 2022; Resende, Melo, Reis, Vasconcelos, Almeida and Benevenuto 2019; de Freitas Melo et al. 2019) or using nationally representative surveys to understand social media habits, information consumption, and exposure to misinformation on WhatsApp (Rossini, Baptista, de Oliveira and Stromer-Galley 2021; Gil de Zúñiga, Ardèvol-Abreu and Casero-Ripollés 2021). The few experimental studies on WhatsApp focus on exposure to misinformation corrections using real or simulated WhatsApp environments (Badrinathan and Chauchard 2023; Bowles et al. 2025), but are limited in making claims about how direct WhatsApp usage influences exposure to misinformation, belief in misinformation, and political attitudes. Through a multi-country field experiment, our study adds to this body of research with causal evidence on the effects of WhatsApp usage on exposure to various types of information and how such exposure is related to attitudes.

2 Theory

How social media platforms affect citizens' political attitudes and behaviors is often associated with how these platforms' affordances potentially enable the spread of low-quality, polarizing

or purely false content (Guess, Nyhan and Reifler 2018; Flaxman, Goel and Rao 2016; Aruguete, Calvo and Ventura 2023). Such concerns are particularly pronounced during periods of high political polarization, including election seasons. A rapidly expanding body of research, from a diverse range of disciplines, has explored the broad societal implications of social media use, including its role in exacerbating political polarization (Banks et al. 2021; Tokita, Guess and Tarnita 2021; Bail et al. 2018; Settle 2018), distorting beliefs about the veracity of information (Pennycook, Cannon and Rand 2018; Anspach and Carlson 2020), deepening policy divisions (Velez and Liu 2024), and negatively affecting mental health, especially among younger users (Tromholt 2016; Vanman, Baker and Tobin 2018; Hanley, Watt and Coventry 2019).

Yet measuring the causal effects of social media usage on what users consume online and its downstream attitudes remains a significant methodological challenge. In the absence of natural experiments varying early adoption of these tools, the population of interest is, in most cases, already heavy consumers of scholars' treatments of interest. Facing these challenges, researchers have used deactivation studies to understand the informational, political, and non-political effects of social media usage (Allcott et al. 2020, 2024; Asimovic et al. 2021; Asimovic, Nagler and Tucker 2023; Arceneaux et al. 2023). Our paper uses a similar design, incentivizing participants in three major Global South countries to reduce their usage of the most popular social media and messaging app in the world for a four-week period before elections.

In this section, we first theorize about the pre-registered first-stage effects of our interventions, focusing on how our intervention might affect participants' informational environment, particularly the consumption of misinformation and low-quality online content. We then explore how these changes propagate to political attitudes and then to non-political attitudes and substitutes for social media usage.

2.1 Consequences for Online Information Consumption

We posit that reducing usage on platforms like WhatsApp—while still allowing individuals to use these platforms for their daily communication needs—can effectively change participants' informational environment. Specifically, usage reduction may reduce exposure to misinformation and low-quality online content through two primary mechanisms: (1) greater attention paid to the

content one is viewing, and (2) prioritization of what is being viewed in the first place.

First, existing research suggests that misinformation often takes hold when individuals are not fully engaged with the content they consume (Pennycook et al. 2021). On platforms like WhatsApp, where misinformation frequently spreads anonymously through viral multimedia content—and specifically, primarily through images (Resende, Melo, Sousa, Messias, Vasconcelos, Almeida and Benevenuto 2019)—simply raising the barriers to passive consumption may lead to significant changes in the type of information people consume. For example, a subtle change like turning off automatic multimedia downloads—effectively adding friction to the first visualization of media content on WhatsApp—requires users to make an active choice to view content, thereby increasing their likelihood of both paying attention to (1) whether they wish to view the content and (2) the content's information itself. This small adjustment could substantially reduce the passive intake of misinformation, as it nudges users toward a more deliberate consumption of media.

Second, a simple reduction in consumption across the board, similar to other social media deactivation designs, can play a crucial role. News consumption on social media platforms tend to be incidental through large groups, meaning that people do not prioritize seeking out news directly on a daily basis (Boczkowski, Mitchelstein and Matassi 2018; Masip et al. 2021). Thus, if participants are prompted to reduce WhatsApp usage, we can reasonably expect that they would likely focus their limited time towards connecting with close networks (such as friends and family) or focusing on work-related communications. Such prioritization should lead participants to reduce the amount of time they spend on consuming content from larger, anonymous groups or channels, where the quality of information is more likely to be low, or about broader socio-political matters that might not affect them in that moment. By reducing and thus curating their consumption habits, users may become more likely to engage with reliable information from trusted sources, effectively filtering out much of the noise and misinformation that thrives in less personal, more public spaces.

In short, both raising the stakes for multimedia consumption and encouraging selective engagement through overall consumption reduction can create an environment where low-quality content, including misinformation, struggles to gain a foothold, leading to a more informed and discerning user base. Yet, prioritizing certain forms of engagement with social media may also re-

duce the consumption of *truthful* news published during the electoral cycle. Because participants are making determinations about what to consume based on contents' attributes (multimedia) and sources (close networks versus larger channels), it may be difficult to discriminate between different types of content that are or are not being selected for consumption. Thus, we hypothesize the following:

H1a: Reducing WhatsApp usage reduces exposure to *false* information.
H1b: Reducing WhatsApp usage reduces exposure to *true* news.

Reducing social media usage is not just about limiting the flow of information; it can also fundamentally alter how individuals form and solidify their beliefs. Existing research has demonstrated that repeated exposure to information increases the likelihood of believing it is true, irrespective of its accuracy (Pennycook, Cannon and Rand 2018). For example, once misinformation is consumed, false beliefs may be difficult to correct due to motivated reasoning (Ecker et al. 2022; Bradley, Angelini and Lee 2007; Martel, Pennycook and Rand 2020). This finding underscores the importance of consumption patterns in shaping public opinion: when individuals are exposed to less content, they might be more skeptical and less confident in the veracity of the information they encounter. This reduced certainty impacts the acceptance of both false and true information, making it harder for misinformation to take root but also introducing doubt in truthful news circulating on social media. We therefore hypothesize the following:

H2a: Reducing WhatsApp usage increases the likelihood of accurately identifying *false* information as false.
H2b: Reducing WhatsApp usage decreases the likelihood of accurately identifying *true* information as true.

Beyond reducing the volume of information people are exposed to before elections, decreasing WhatsApp usage can also shape the type of political and non-political content individuals recall, particularly mitigating exposure to uncivil and intolerant speech (Chen 2017; Rossini 2022), which may often also feature misinformation or otherwise politically polarizing content (Anspach and Carlson 2020). This follows a similar logic as above: as users prioritize interactions within close-knit networks or focus on work-related communications, their engagement with larger, impersonal channels or large group chats ought to diminish. These broader forums are often breeding

grounds for toxic discourse and the sharing of inflammatory content as they often amplify the most extreme and uncivil voices (Bor and Petersen 2022), while online anonymity and distance encourage negative behaviors that are less likely to occur in more personal settings (Rowe 2015). Overall, by curbing time spent in these spaces, users are not only avoiding misinformation but are also distancing themselves from the negativity that can pervade online discussions. We therefore hypothesize:

H3: Reducing WhatsApp usage decreases exposure to low-quality political discussions.
H4: Reducing WhatsApp usage decreases exposure to uncivil political content online.

2.2 Downstream Consequences on Political Attitudes

Thus far, we have argued that reducing WhatsApp usage—through either reduced consumption of multimedia or through reduced time spent on the platform—ought to shift users' (1) exposure to information, (2) perceived accuracy of information, and (3) exposure to broader negative political discourse. We next investigate how these direct effects may have downstream impacts on salient political and social outcomes.

We first explore downstream consequences on political attitudes, and especially political polarization, which is often cited as a negative consequence of social media usage and broader internet access (Allcott et al. 2020; Arugute, Calvo and Ventura 2022; Bail et al. 2018; Lelkes, Sood and Iyengar 2017; Settle 2018). Polarization is important because it shapes not only political but also social and economic spheres: when societies are deeply divided along social or ideological lines, and when politics becomes factionalized, policy disagreements can escalate into negative perceptions of the opposition and social alienation of outgroup members (Iyengar et al. 2019). This, in turn, can weaken social cohesion, erode democratic norms, exacerbate policy gridlock, and even fuel political violence (Kingzette et al. 2021; Svolik 2019; Piazza 2023; Badrinathan, Chauchard and Siddiqui 2024). Political polarization may pertain to either social (*affective*) or ideological (*issue*) divisions. We highlight that these two forms of polarization are rooted in distinct socio-political identities: affective polarization, where political partisans view opposing groups with hostility (Iyengar, Sood and Lelkes 2012), and identity-based prejudice rooted in ethnic or racial divisions (Horowitz 1993; Wilkinson 2006).

We posit that reducing WhatsApp usage—together with the first-order effects described in the previous hypotheses—may affect users' attitudes towards other political and social groups in society through distinct mechanisms. First, during elections, information (true and false) tend to be produced and consumed across partisan lines, often associated with outgroup negativity (Rathje, Van Bavel and Van Der Linden 2021), which can exacerbate how hostile people feel towards members of opposing political and social groups (Kingzette et al. 2021; Jenke 2024). Second, research has shown that politically rooted misinformation and misperceptions about outgroups (for example, overestimating how willing a particular group is to not accept the results of elections) play a key role in heightening levels of polarization (Druckman et al. 2023; Voelkel et al. 2023). These misperceptions are tightly connected to how social media platforms give voice to a small number of users who post only their most extreme opinions and do so at a very high volume (Aruguete, Calvo and Ventura 2023; Osmundsen et al. 2021). Third, these polarizing effects are not restricted to misinformation consumption. Recent work has shown that when voters have their preferences confronted with contentious, uncivil arguments—as descriptive work argues is common in group setting dynamics on WhatsApp (Chauchard and Garimella 2022)—voters tend to double-down on their preferences, and polarization tends to increase (Velez and Liu 2024). Furthermore, existing research shows that interventions aimed at reducing exposure to misinformation and polarizing content, such as fact-checking corrections (Druckman et al. 2023), counter-attitudinal media exposure (Levy 2021), or temporary deactivation from social media platforms like Facebook (Allcott et al. 2020, 2024), may mitigate affective polarization.[3] Thus, in the context of our study, we hypothesize:

H5: Reducing WhatsApp usage reduces affective partisan polarization.
H6: Reducing WhatsApp usage reduces identity-based outgroup prejudice.

Reducing WhatsApp usage may also reduce polarization directly related to the election itself. We consider two forms of ideological polarization: issue polarization, where people become increasingly divided over specific political topics and party platforms, and candidate favorability, where people strongly prefer one candidate over others. Decreased engagement with negative political discourse, particularly with relation to ongoing discussions about policy proposals and

[3] Although note that Asimovic et al. (2021) found that deactivating from Facebook at an politically salient period of time in Bosnia-Hercegovina actually *exacerbated* affective polarization from an ethinc out-group attitude perspective.

party campaigning, reduces the likelihood that people are polarized by partisan media coverage, which might lead them to better consider issues themselves (Levendusky 2013; Velez and Liu 2024). More mechanically, through reduced exposure to *true* news, people may also be less certain about the issues altogether. In the same vein, reducing exposure to biased or inflammatory content during the election period might also temper the extremes of candidate favorability: by seeing less in- vs. out-group rhetoric, people ought to be less likely to intensely favor their preferred party's candidate(s) and oppose the other party's candidate(s). These effects are critical, particularly in the Global South context where social media has become the primary way through which many voters report receiving news and learning about elections (Newman et al. 2024), and where political elites rely heavily on social media to communicate with their constituents (Bessone et al. 2022; Wirtschafter et al. 2024). We therefore hypothesize:

H7: Reducing WhatsApp usage reduces issue polarization.
H8: Reducing WhatsApp usage reduces relative candidate favorability.

2.3 Downstream Consequences on Substitutes and Non-Political Attitudes

Lastly, we examine the downstream effects of reducing WhatsApp usage on outcomes that are not explicitly political but help provide a more comprehensive picture of our intervention. In particular, we assess how reducing WhatsApp usage can lead to (1) substitution to different online and offline activities and (2) changes in subjective well-being. First, to the extent that treated participants reallocate time previously spent on WhatsApp to other activities, our estimated effects reflect both the direct impact of reduced WhatsApp use as well as indirect substitution dynamics. Therefore, we analyze self-reported substitution patterns in detail, with particular attention to shifts between online and offline behaviors. Second, the growing role of social media in daily life has sparked important debates about its consequences for subjective well-being. Prior research highlights several potential mechanisms for how social media usage might reduce subjective well-being — such as reduced face-to-face interaction, increased social comparison (Kross et al. 2021; Twenge and Campbell 2018), and information overload (Matthes et al. 2020; Goyanes, Ardèvol-Abreu and Gil de Zúñiga 2023)We note that our analyses of these outcomes are exploratory, as they were pre-registered as additional research questions without directional hypotheses. Both

subjective well-being and substitutes have been incorporated in previous social media deactivation experiments.

3 Experimental design

In 2024, we implemented a multi-country field experiment during the four weeks preceding elections in three major Global South democracies: India and South Africa (which held general elections) in the spring and Brazil (which held municipal elections) in the fall. In India, polling occurred between April 19 and June 1; results were announced on June 4. In South Africa, the general election was held on May 29, and results were announced shortly thereafter on June 2. Given the similar timelines, our usage reduction experiment occurred simultaneously in these two countries. We then repeated our experiment in Brazil, which had local elections for all municipalities and local councils on October 6. Our design randomly assigned half of the participants to change their WhatsApp usage, either through a multimedia deactivation or a screen time reduction, for a period of four weeks (that is, from April 29 to May 27 in India and South Africa, and from September 9 to October 4 in Brazil). We then administered a final survey to measure the effects of participating in the experiment. The research received approval from the Institutional Review Boards at [REDACTED] In this section, we describe each stage of the study in further detail.

3.1 Recruitment

We recruited participants for the study using the Meta Advertisements platform. Our ads were posted on Facebook, Instagram, and Messenger (see SM Section A.1 for advertisement examples) and linked participants to a short Qualtrics survey (*Recruitment Survey*). In this survey, potential participants provided consent and answered questions pertinent to study eligibility and block randomization. These include questions about respondents' demographics, their social media habits (particularly focusing on WhatsApp usage), and their willingness to join a four-week study that potentially involves reducing WhatsApp usage. The survey also collected contact information (email address and WhatsApp phone number) so that we could follow up about the experiment afterwards.

Using responses from the recruitment survey, we applied the following criteria to determine

eligibility:

- Participants should be 18 years or older;
- Participants should report using WhatsApp for at least 10 minutes per day;
- Participants should not report using WhatsApp on a shared device;
- Participants should have spent at least one minute completing the recruitment survey;
- Participants should not be identified as bots, spammers, or duplicate observations, as determined by Qualtrics security filters and contact information.

In total, our recruitment and filtering procedure yielded 1,310 eligible participants in India, 2,884 in South Africa, and 2,067 in Brazil. All of the surveys described below carried monetary incentives. SM Section A.3 details our incentive structure, which was communicated to participants in general terms in the recruitment advertisement and survey, and again with more elaboration at the end of the baseline survey.

3.2 Baseline & Treatment Assignment

We invited all eligible participants to join our study through email and WhatsApp messages. Participants were taken to a Qualtrics survey (*Baseline Survey*), which had three components. First, we asked a set of questions about pre-treatment covariates of interest and collected pre-treatment measurements of some of our eventual outcomes. Second, participants provided their baseline WhatsApp usage information by submitting screenshots of their WhatsApp settings page (explained further below). Third, in the last stage of the survey, we informed participants of their pre-randomized treatment condition. Prior to fielding the baseline survey, participants were block-randomized on age, education, gender and self-reported WhatsApp usage and then assigned to one of four equal-sized groups, listed below. In total, the number of people who successfully enrolled in the experiment was 678 (52%) in India, 820 (28.5%) in South Africa, and 928 (44%) in Brazil.[4] In SM Section C.1, we test for baseline differences in pre-treatment covariates between participants invited and participants enrolled in the experiment.

- **Time:** Set the WhatsApp app timer to ten minutes AND reduce WhatsApp usage to ten minutes per day for four weeks

[4]The vast majority of the people contacted via email and WhatsApp did not started the survey. The share of people successfully enrolled in the experiment *conditional* on the participants who at least responded to the first survey question of the treatment assignment survey is 83.3% in Brazil, 90% in India, and 77% in South Africa, for a total of 1111 participants in Brazil, 755 in India, and 1056 in South Africa.

- **Control (Time):** Reduce WhatsApp usage to ten minutes per day for three days.
- **Multimedia:** Turn off the automatic download of media on WhatsApp AND do not consume any media on WhatsApp for four weeks.
- **Control (Multimedia):** Do not consume any media on WhatsApp for three days.

To facilitate compliance, participants in the two treatment groups were instructed to make a change to their WhatsApp settings. Participants assigned to **Time** were asked to set their WhatsApp app timer to 10 minutes per day, while participants assigned to **Multimedia** were asked to turn off their automatic download of audio, images, documents, and video on WhatsApp. Neither of these interventions imposed "hard constraints" on participants' WhatsApp experience. Instead, these modifications worked as nudges, adding barriers to participants' regular WhatsApp usage. In the **Time** condition, the app timer only reminded participants when they passed their screen time daily limit, requiring them to leave the app or override the option for a few more minutes. In the **Multimedia** condition, turning off automatic downloads on WhatsApp introduced substantial friction to consuming content by blurring all multimedia received in personal conversations and group-based chats. This setting required users to undertake an extra tap to view the content; users in our treatment condition were asked to refrain from doing that. Table 1 presents a summary of the instructions for each group.

To avoid one potential source of differential attrition, our experimental design only informed the participants about their assignment (control or treatment) after they uploaded their first screenshot showing their baseline time/media usage (SM Section A.2 presents examples of the screenshots we required participants to submit). For the same reason, we asked the control group to spend three days performing the same task asked of the treatment group. The purpose of requesting the control groups to undertake the same task as the treatment groups for a shorter period of time was to ensure that all participants were willing and able to join the experiment. However, while participants were nominally "treated" for a brief period, we expect no effects from this short, suggestive intervention to persist until the endline survey (four weeks later).

3.3 Compliance Tasks

We monitored compliance by asking respondents to provide us with screenshots from their mobile devices showing their WhatsApp time or media usage statistics, depending on what their

Table 1: Summary of Instructions for Treatment and Control Groups

Condition	Instructions & Steps
Time	i) Submit a screenshot showing their WhatsApp screen time usage for the past week (Figures A6A and B). ii) Informed of treatment: Limit WhatsApp usage to 10 minutes per day for four weeks. iii) Given instructions on how to set a 10min daily timer on their WhatsApp app. iv) Submit a screenshot showing that the app timer has been added (Figure A7).
Control (Time)	i) Submit a screenshot showing their WhatsApp screen time usage for the past week (Figures A6A and B). ii) Informed of treatment: Limit WhatsApp usage to 10 minutes per day for three days.
Multimedia	i) Submit a screenshot showing their WhatsApp media storage statistics (Figures A6C and D). ii) Informed of treatment: Do not consume any multimedia content on WhatsApp for four weeks. iii) Given instructions on how to turn off the automatic download of media on WhatsApp. iv) Submit a screenshot showing that the automatic download of media has been turned off (Figure A7C and D).
Control (Multimedia)	i) Submit a screenshot showing their WhatsApp media storage statistics (Figures A6C and D). ii) Informed of treatment: Do not consume any multimedia content on WhatsApp for three days.

treatment assignment was. Once a week, we sent participants a short Qualtrics survey, prompting them to submit these screenshots. In total, participants submitted four such screenshots, in addition to the initial screenshot collected during the baseline survey.

For participants in the Time conditions, we asked them to send us screenshots of their daily usage of WhatsApp covering every week of the experiment, as in Figures A6A and B. This information is easily available in the Settings app of most mobile devices as part of digital well-being

monitoring tools. For participants in the media condition, we asked them to send us screenshots of their WhatsApp storage information for media received, as in Figure A6C and D. This page, which is a part of WhatsApp and therefore accessible on all devices, records the volume of bytes of media received through WhatsApp.

3.4 Post-Treatment Survey

After the four-week experiment, we invited respondents to answer a final survey. The outcomes collected through this survey are described in detail in the next section, where we present our findings regarding the effects of reducing WhatsApp usage during elections (see SM Section B for a summarized presentation). The survey was sent to Indian and South African participants on May 27, a few days before election results were announced in their countries, and to Brazilian participants on October 4, three days before the election. We chose to send the surveys before the announcement of results in India and South Africa in order to avoid the potential risks of restricting participants' access to WhatsApp post-election. We decided not to prevent them from receiving news about the election outcomes and potential offline events related to the elections, such as episodes of violence, riots, or strikes. In Brazil, the results are announced hours after polls are closed. In total, the number of people who completed the post-treatment survey was 653 (96%) in India, 742 (90%) in South Africa, and 825 (89%) in Brazil.

In multi-wave field experiments, a primary concern for internal validity is the presence of differential attrition between the treatment and control groups. SM Section C.2 thoroughly investigates this issue. Combining data from all three countries, 2,220 out of 2,425 enrolled participants did not complete the post-treatment survey (with an attrition rate of 9%). There is no evidence of differential attrition between treatment and control (β=0.002, $p-$value=0.88; see SM Table C7). Following the attrition bias test suggested in Ghanem, Hirshleifer and Ortiz-Becerra (2023), we examine the presence of selective attrition, by comparing baseline characteristics in pre-treatment covariates and outcomes between attriters and completers in the treatment and control. We find no evidence of selective attrition affecting the internal validity of our design. In other words, treated and control attriters are not jointly distinguishable from participants who completed the study.

4 Results

In this section, we start by examining the degree to which treated users actually reduced WhatsApp usage relative to users in the control condition (Section4.1). We then present the results of our pre-registered analyses, first considering outcomes related to information outcomes (Section 4.2) and then outcomes related to broader social and political attitudes (Section 4.3). Finally, we present additional analyses related to non-political outcomes (Section 4.4).

In our regression analyses, we estimate the adjusted intention-to-treat (ITT) effect of the corresponding treatment on our outcomes of interest using OLS estimators (see SM D.1 for unadjusted ITT results). As pre-registered, we add as covariates the variables used in our block randomization procedure (age, gender, education, and self-reported WhatsApp usage) and additional pre-treatment variables selected via Lasso for each outcome. Our primary specification pools the two treatment arms across all countries (see SM Section E.1 for unpooled treatment arm effects [5]; for each outcome, we also present pooled treatment effects by country and unpooled country-level effects. To account for the distinct baselines of the outcomes in each country, all models use a multilevel estimation with random intercepts at the country level. All confidence intervals use a two-sided test with $p < 0.05$ as our measure of statistical significance. In the supplemental materials (SM E.5), we present results with multiple hypotheses adjustments.

4.1 Compliance with the WhatsApp Usage Reduction Intervention

We incentivized users in the treatment condition to alter their WhatsApp usage, either by limiting the amount of time they spend on the app or by refraining from consuming multimedia content on the app. Did this actually lead to reduced WhatsApp usage/consumption? To answer this question, at the end of every week of the experiment period, we collected a screenshot showing one's weekly WhatsApp screen time from participants in the Time condition, and a screenshot showing one's cumulative WhatsApp data consumption from participants in the Media condition. We use these screenshots to construct a binary indicator of "low WhatsApp usage" which serves

[5] In the unpooled models, to increase statistical power, we compare each treatment arm with a pooled control group. Because both control groups resumed their regular WhatsApp activity after three days, we do not expect any persistent effect at the time of the post-treatment survey. The SM section E.7 shows that all primary outcomes have non-distinguishable mean differences when comparing the media and time control groups in the post-treatment survey

as an indicator of compliance for those in the treatment condition and a benchmark to compare against that in the control condition. A user in the Time condition is classified as having low WhatsApp usage if each of their last three weekly screenshots show a daily average WhatsApp screen time of less than 10 minutes. A user in the Media treatment condition is classified as having low WhatsApp usage if the gap in cumulative megabytes downloaded between their baseline screenshot and endline screenshot is lower than the average gap in the control condition in their respective country; a user in the Media control condition is classified as such if this gap is lower than 100 megabytes, which we determined to be a reasonable threshold following pilot studies.[6] The classification criteria in both arms are the strictest possible criteria, and therefore produce the lowest bounds of compliance in the treatment conditions.

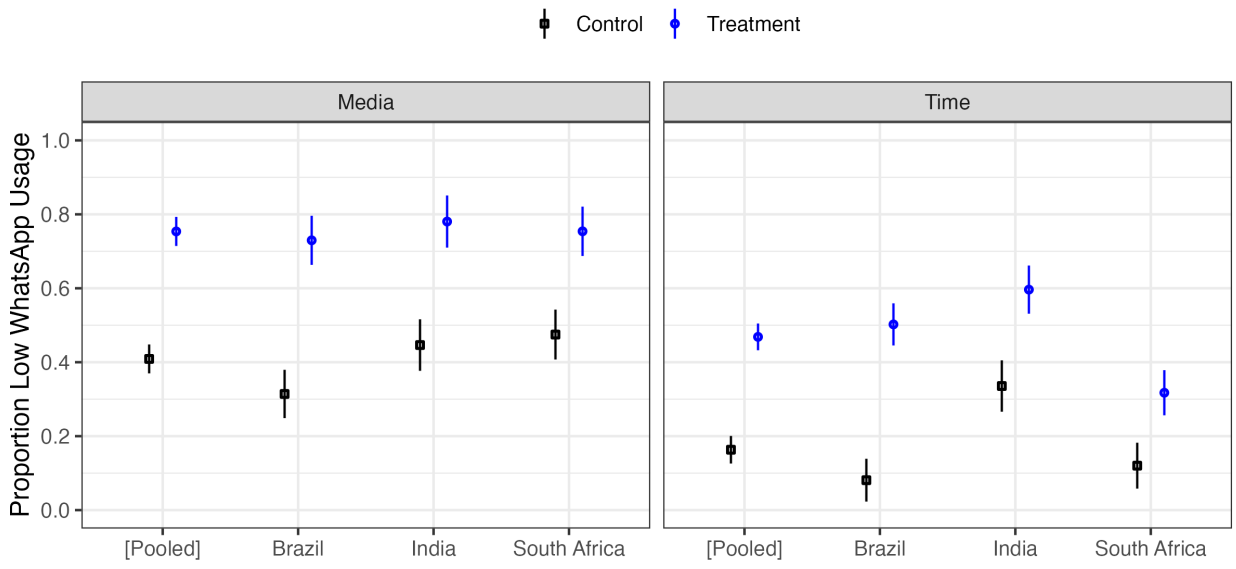
Figure 1 shows the proportion of treatment (in blue) and control (in black) participants classified as having low WhatsApp usage in the Media arm (first panel) and in the Time arm (second panel). First, we observe considerable variation in control group usage levels across countries, with Brazil having relatively fewer low usage respondents (and relatedly larger treatment effects on reducing usage). Second, across countries and arms, the proportion of participants with low WhatsApp usage is significantly higher in the treatment condition than in the control condition, showing that treatment assignment worked as intended in creating two distinct groups with substantially different levels of time spent on WhatsApp or multimedia content consumed during elections. Third, we note that under the strictest definitions of compliance, there is a stronger degree of compliance in the Media arm compared to the Time arm, suggesting that, given the same monetary incentives, duration, and baseline usage levels, individuals are more willing and able to stop consuming multimedia content on WhatsApp compared to limiting WhatsApp usage overall to 10 minutes per day. This is an important descriptive finding for future academic research and policymaking on WhatsApp, confirming our priors that a full WhatsApp deactivation is infeasible in settings where it is much more embedded in users' lives relative to social media platforms like Facebook would be.

That said, though we observe that the proportion of Time treatment respondents who spent

[6] If a user submits a screenshot that is doctored, is a duplicate of a previous week's submission, is from a different phone as their previous submissions, or indicates their data consumption statistics have been reset during the experiment period, we classify them as having violated study rules and not having met our low WhatsApp usage criteria. In practice, about 4% of our sample violated one of these rules; country or treatment status does not predict violations.

the *entire* experiment period under the 10-minute daily limit was about 0.47 (as shown in Figure 1), the proportion who never exceeded 30 minutes was over 0.70. Further, in any given week, the proportion who stayed under 10 minutes was over 0.60. Comparing usage in the pre-treatment week with the average during-treatment week, we also observe a reduction of about 0.84 SD in WhatsApp screen time within the treatment condition. These numbers confirm that the Time treatment was quite effective in spurring respondents to limit their usage.

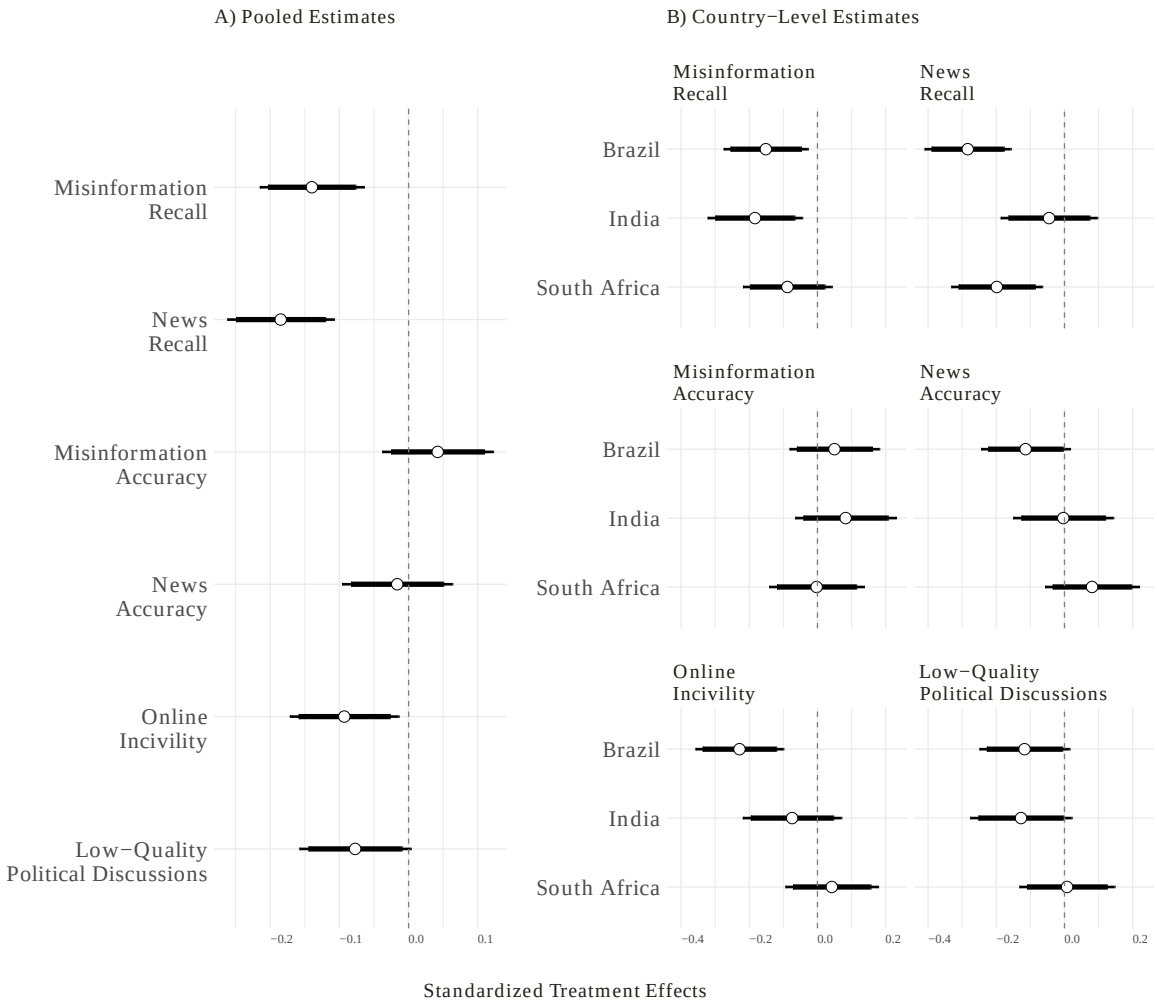
Figure 1: Treatment Assignment and Low WhatsApp Usage During Experiment



4.2 The Effects of WhatsApp Usage on Types of (Mis)Information Consumed

In Figure 2, we present the effects of reducing WhatsApp usage on outcomes related to exposure to different types of information and discourse on WhatsApp. We hypothesized that our intervention should first reduce exposure to both false and true information (**H1a**, **H1b**). These effects on exposure should then increase (decrease) treated participants' ability to identify false (true) information (**H2a**, **H2b**). We also expected reduced exposure to low-quality political discussions (**H3**) and incivility online (**H4)**.

Figure 2: Treatment Effects on Information Outcomes



4.2.1 News and misinformation: exposure and belief

We first test treatment effects on exposure to and belief in false information and true news. In our endline survey, we provided participants with two distinct sets of headlines—*News Headlines* and *Misinformation Rumors*—to assess how reducing WhatsApp usage affects participants' exposure to political (mis)information stories and their beliefs about these stories.

To identify pertinent *News Headlines*, we selected six stories that appeared on major news organizations' websites in each country during the four weeks of the experiment.[7] We further divided

[7]To identity salient news headlines, we scraped Google News daily throughout the four-week experiment period and selected a representative sample of six stories covering different issue areas, partisan groups, and time points in the

these six headlines into *True News* and *Placebo False News*. *True News* comprises four headlines that appeared as is in newspapers, while *Placebo False News* were altered versions of two distinct true news headlines that appeared in the media. These modifications introduced a false fact—usually the reverse of the original true news—to avoid acquiescence bias for the News Headlines outcome (Hill and Roberts 2023). This procedure to select News headlines follows other recent deactivation studies (Allcott et al. 2020, 2024; Asimovic et al. 2021; Asimovic, Nagler and Tucker 2023; Arceneaux et al. 2023; Ventura et al. 2025).

Meanwhile, *Misinformation Rumors* comprises four false stories that circulated widely on social media during the election. While the most ideal measure would use the most salient stories circulating on WhatsApp during this period, there is no straightforward or reliable way to track such data due to the platform's end-to-end encryption. As an alternative, we instead reviewed major, well-reputed fact-checking websites in each country and included stories that were checked by most agencies during these four weeks. As with *True News*, we selected a diverse set of misinformation headlines corresponding to different topics, political leanings, and weeks of the experimental period. In the SM Section B.1, we provide the full text of all headlines included in the three countries.

For each headline, we asked participants (a) whether they have seen this headline in the past 30 days and (b) the extent to which they think the headline is accurate. Our final pre-registered outcomes, presented in Figure 2, are operationalized as follows. First, `Misinformation Recall` counts the number of misinformation rumors (out of four) that the participant has seen in the past 30 days; similarly, `News Recall` counts the number of true news stories (out of four) recalled by each participant. We drop the two placebo headlines when estimating treatment effects on information exposure, since these were headlines that were artificially modified by us and which, by definition, respondents could not have seen. Second, for information beliefs, we count the number of stories correctly identified as true or false. Thus, `Misinformation Accuracy` ranges from 0 to 4. `News Accuracy` range from 0 to 6, as we include the placebo headlines in our conceptualization of news accuracy.

The first and second outcomes in Figure 2(A) show that, consistent with **H1a**, limiting WhatsApp usage—be it by maintaining time limits or by refraining from consuming multimedia content—

experiment.

reduced recall of political misinformation rumors circulating widely in the weeks of the intervention. The intention-to-treat analysis shows a reduction in exposure to misinformation rumors of 0.14 SD (p-value < 0.01). In support of **H1b**, we also find a 0.19 SD (p-value < 0.01) decrease in recall of true news stories. While the overall effect sizes are comparable across misinformation stories and true news stories, we note from Figure 2(B) that the reduction in misinformation recall is consistent across the three countries, while the decline in true news recall is observed predominantly in South Africa and Brazil. The different finding in India may be driven by different news consumption patterns: in Appendix Figure E17, we find among the Control group that news exposure in India is higher than in South Africa and Brazil, and thus participants may have more outlets through which they were able to stay informed about current events. These results are not affected by adjustments for multiple comparisons (see SM Section E.5).

Turning to perceptions of accuracy, the third and fourth outcomes in Figure 2(A) present estimates corresponding to belief in misinformation stories and knowledge of true news stories. We find limited support for **H2a** and **H2b**: although the treatments significantly reduced exposure to both true and false headlines, they do not have a statistically significant effect on belief accuracy for either. These null effects align with previous deactivation studies, which have also failed to shift accuracy perceptions of false information (Allcott et al. 2020; Ventura et al. 2025) or general news knowledge (Allcott et al. 2024). In sum, changes in self-reported exposure do not directly translate to updates in belief accuracy, suggesting a potentially more complex mapping from exposure to accuracy belief formation than posited in extant literature.

In SM Figure E8, we present unpooled treatment effects and show that the two treatment arms produced similar effects—both in effect size and statistical significance—across the information recall and accuracy outcomes. Thus, both types of WhatsApp usage reduction similarly decreased exposure to misinformation and true news. Notably, in our survey, we separately asked participants to recall how much true and false news they saw on social media over the past month. In SM Figure E11, we show that participants reported seeing less false news on social media but not true news. In other words, although actual consumption of both types of information declined due to treatment, treated participants were only cognizant of reductions in misinformation exposure. This asymmetry suggests that individuals may be less attuned to their consumption of true news on social media more broadly.

4.2.2 Online Incivility and Quality of Political Discussions

Beyond measuring direct recall and beliefs vis-à-vis true and false information, we are also interested in measuring changes in participants' informational environment during the one-month intervention. To that end, we focus on two primary outcomes: exposure to toxic content online and to low-quality political discussions online and offline. First, to measure participants' exposure to online toxicity and incivility, we ask them if, during the past month, they were targeted with hostile comments online or saw online comments that were rude or disrespectful. We build an index with these two measures called *Online Incivility*. Second, to measure the overall effects of our intervention on participants' broader engagement with political discussions, we use a battery of items asking participants about their exposure to discussions that made them angry (Allcott et al. 2020), that they consider uncivil (Rossini 2022; Chen 2017), and that helped them reduce prejudice towards outgroups (Allcott et al. 2020). We build an index with these three measures called *Low-Quality Political Discussions*.

Figure 2 shows that the reduction in WhatsApp usage rendered a significant reduction in self-reported exposure to online toxic speech and exposure to low-quality discussions about politics, consistent with **H4** and **H5**. The intention-to-treat analysis shows a reduction of 0.09 SD (p-value < 0.01) in exposure to online toxicity and of 0.07 SD (p-value < 0.10) in exposure to low-quality political discussion. For online toxicity, the effects are primarily driven by the Brazilian sample, while for low-quality political discussion, both Brazil and India show consistent reductions. The treatment's particular effectiveness on these outcomes in Brazil may reflect a combination of two factors: (1) Brazilian participants' somewhat higher exposure to incivility online (see Appendix Figure E16 for Control group values), which were reduced as a result of treatment; and (2) their greater tendency to substitute WhatsApp usage with offline activities such as hobbies or time with friends (see Figure 4).

In SM Figure E8, we further disaggregated our two treatment arms to explore their differential effects on these outcomes. We find that reductions in exposure to online incivility are primarily driven by turning off automatic media downloads, whereas reductions vis-à-vis low-quality political discussions are predominantly driven by capping overall WhatsApp screen time. It is possible that reduced exposure to inflammatory media content mitigates perceptions of online exposure to

uncivil content, while minimal engagement with the platform, through a usage reduction, limits participation in toxic political discussions. Generally, these effects echo anecdotal and descriptive evidence of the role of social media in increasing exposure to hostile content and low-quality information about political issues (Bor and Petersen 2022; Recuero, Soares and Vinhas 2021).

4.3 Downstream Consequences: The Effects of WhatsApp Usage on Political Attitudes

We theorized that reducing WhatsApp usage ought to also have downstream consequences for users' political attitudes, by reducing information exposure and altering beliefs, or by limiting exposure to negative discourse more generally. In Figure 3, we show how our treatment arms affected salient political outcomes, namely partisan polarization (**H5**), identity-based outgroup prejudice (**H6)**, issue polarization (**H7**), and relative candidate favorability (**H8)**.

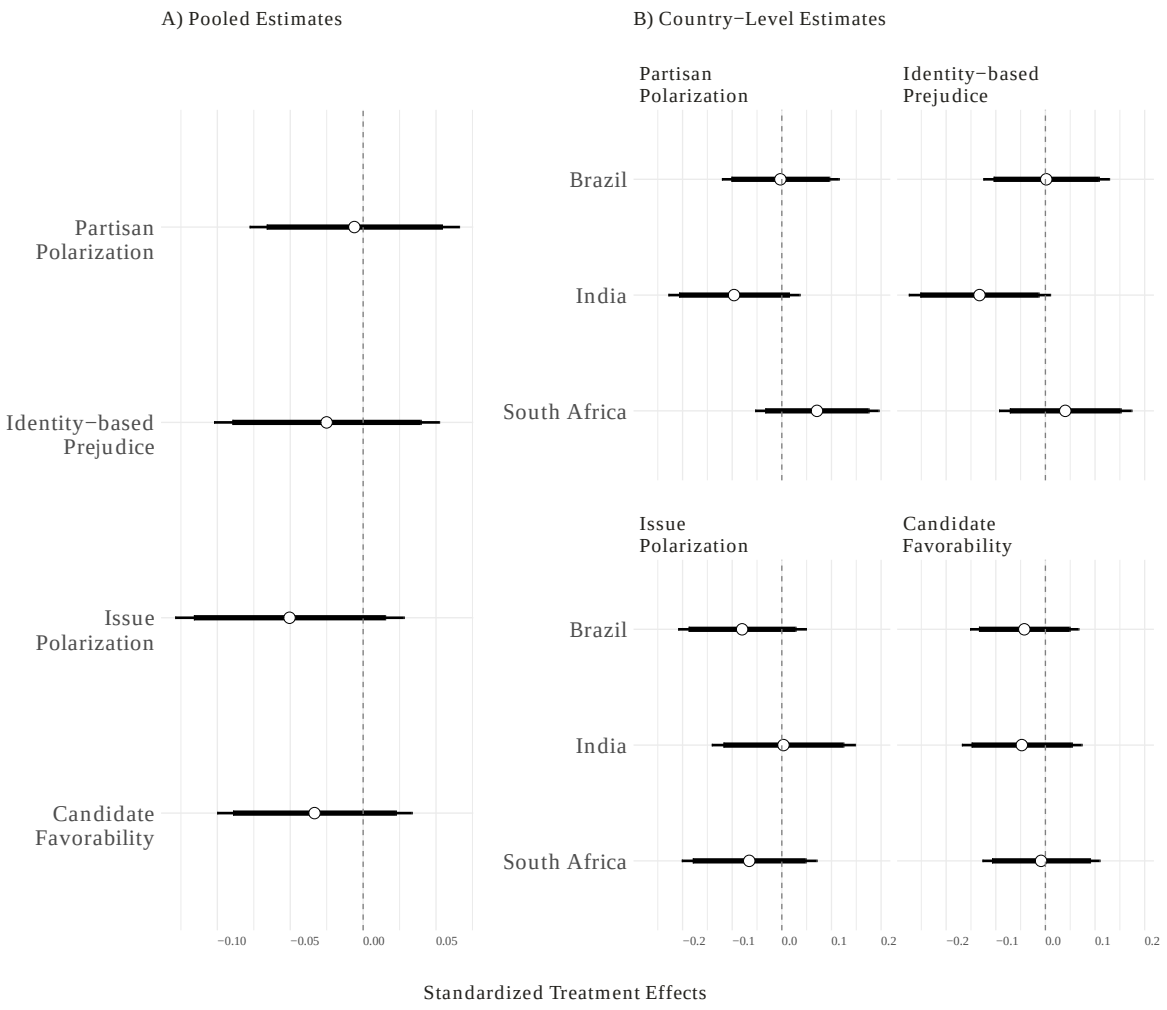
4.3.1 Political Polarization and Identity-Based Prejudice

Reduced exposure to misinformation and campaign content denigrating political parties and their supporters might reduce animosity towards voters of those parties. To test this hypothesis (**H5**), we collected participants' views about voters of the two largest parties in their country, which also represent the main opposing coalitions in the 2024 elections. In India, these two parties are the Bharatiya Janata Party (BJP) and the Indian National Congress (Congress Party); in South Africa, the African National Congress (ANC) and the Democratic Alliance (DA); and in Brazil, the Partido Liberal (PL) and the Partido dos Trabalhadores (PT) .[8] We define each participant's *in-party* as the party representing the coalition they would prefer to see win the election, and their *out-party* as the party leading the other coalition. In our sample of Indians, 68.7% (31.3%) listed the BJP-led coalition (Congress-led coalition) as their preferred choice. In South Africa, 49.1% (50.9%) of our sample preferred the ANC-led coalition (DA-led coalition). In Brazil, 44.2% (55.8%) preferred the PL-led coalition (PT-led coalition).

We measure three outcomes widely used in the literature to create our `Partisan Polarization`

[8]It should be noted that in politically diverse, multi-party systems, many people may *not* identify as voters of one of the two largest parties. We collected extensive data on party preference in our baseline survey and are thus equipped to undertake finer analyses taking into consideration support for different parties beyond coalition support.

Figure 3: Treatment Effects on Political Attitudes



outcome (Iyengar, Sood and Lelkes 2012; Druckman et al. 2021). First, we use a 7-point feelings scale to capture overall feelings towards the two parties' voters. Second, we present a list of social interactions—such as watching a sports game together or living in the same neighborhood—and ask participants which they would be willing to do with voters of each party. Third, we asked participants to read a list of positive and negative traits and indicate which might describe a typical voter of each party. For each of the three items, we calculate the absolute distance between participants' views about these two parties' supporters. We then aggregate these measures into our `Partisan Polarization` index.[9] Thus, for both our index and its three composite items, higher

[9]Table B2 contains more information on our measures.

values represent a larger gap between responses about the two groups. We expect that treated participants would, on average, report smaller gaps than those in in the control group.

The first outcome in Figure 3(A) shows that this expectation does not hold, and thus we find no support for **H5**. Indeed, while our treatments did reduce exposure to partisan misinformation, online incivility and low-quality political content (Figure 2), these changes in the informational environment did not mitigate partisan animosity and polarization. These null results are consistent with previous deactivation studies both on WhatsApp and on other platforms such as Facebook and Instagram (Ventura et al. 2025; Allcott et al. 2024; Arceneaux et al. 2023).

In a similar vein, **H6** predicted a reduction in identity-based (that is, ethnic or racial) out-group prejudice following our detected changes in the participants' informational environment. As outlined in our pre-analysis plan, our `Identity-based Prejudice` analysis focuses only on India and South Africa, as these issues are less central to contemporary Brazilian politics. However, as a validity check, we included and assessed one measure of racial prejudice in our Brazil surveys, which we describe further below.

Political and societal dynamics in India and South Africa are shaped by identity-based prejudice in similar ways, and thus our analysis explores whether reducing WhatsApp usage may reduce intergroup animosity in these two countries. We focus on measuring post-treatment attitudes about the two most politically salient groups in each country—that is, Hindus and Muslims in India and Blacks and Whites in South Africa.[10] In India, there is a history of conflict between Hindus, the dominant religious majority, and Muslims, the largest religious minority. In recent years, WhatsApp has been used to spread misinformation and harmful content that demonizes Muslims. Thus, it is important to examine whether reducing exposure to such content can help to reduce out-group prejudice, especially among Hindus. In South Africa, we examine prejudice along the main racial cleavage (i.e., Black and White South Africans). Legacies of apartheid in South Africa has kept race salient in politics and a frequent subject of misinformation.

Mirroring our partisan polarization estimation, our `Identity-based Prejudice` outcome comprises three items. First, we use a 7-point feelings scale to capture overall feelings towards Hindus

[10]Even though there are many other important ethnic/religious and racial groups in these countries, the majority of information and discourse about ethnic or racial issues relate to Hindus/Muslims in India and Blacks/Whites in South Africa. As such, we did not expect there to be enough content circulating on WhatsApp about any other group such that abstention from WhatsApp would lead to changes in attitudes about said groups. Additionally, to prevent survey fatigue and priming, we refrain from asking questions about all possible ethnic/racial groups in our surveys.

and Muslims in India and towards Blacks and Whites in South Africa. Second, we present participants with a list of social activities and asked them to indicate which ones they would be willing to do with an unknown member of each group. Third, we list a set of traits and ask participants to identify those that might describe a typical member of each group. We then construct our prejudice index by aggregating responses to these three items, where higher values represent a larger gap between responses about the two groups.[11] In Brazil, we only included the 7-point feelings scale, asking respondents to rate their feelings about White and Black Brazilians. Thus, our "index" measure in Brazil only consists of this item.

The second outcome in Figure 3(A) presents treatment effects on identity-based prejudice.[12] Pooling across countries and treatment types yields null results, but obscures an important within-country finding. While we do not uncover treatment effects in South Africa and Brazil, Figure 3(B) shows that there is a marginally significant reduction in ethnic prejudice in India ($d = -0.13$, p-value $= 0.067$). These suggestive findings in India may reflect the fact that Indian participants reported greater interest in politics at baseline and somewhat more trusting of what they see on social media (Appendix Figure E17)—which likely makes them more attentive to political content on social media. Our treatment may have therefore better mitigated this exposure and produced the observed reduction in prejudice.

4.3.2 Issue Polarization and Candidate Favorability

In addition to traditional measures of affective polarization and identity-based prejudice, we explore how WhatsApp usage reduction affects polarization vis-à-vis election-related issues and how voters evaluate political candidates. In doing so, we build on other deactivation studies which have examined similar outcomes (Allcott et al. 2020, 2024).

For issue polarization, we elicit opinions about six different statements related to issues that received public attention during the election period. These issues feature themes of democratic norms, intergroup relations, immigration policy, and recent news in each country. Using these six issues, our `Issue Polarization` index uses the sum of standardized absolute differences

[11] SM Table B3 contains more information on our measures.

[12] It should be noted that our samples are quite skewed in both countries, which is to be expected given the composition of the population in each country. In India, 78.6% of our sample are Hindu. In South Africa, 81.3% are Black. The results are, therefore, predominantly driven by Hindus in India and Blacks in South Africa.

between participants' agreement (on each issue) to the mean issue agreement among the entire control group.[13] Positive values indicate stronger issue polarization, with voters moving away from the average baseline issue agreement, and negative values mean voters move towards the average voter.

For candidate favorability, we use standard a feelings scale featuring major political leaders in each country, including the executive branch incumbent.[14] Our measure is intended to capture how reducing WhatsApp usage might make voters more favorable towards candidates from their preferred party and less favorabe towards alternatives. Therefore, our `Candidate Favorability` index takes the absolute difference between participants' main political ingroup candidate and the candidates from the alternative parties.[15]

The third and fourth outcomes in Figure 3(A) present the pooled effects of our intervention on issue polarization and candidate favorability respectively. Overall, the effects of reducing WhatsApp usage on both outcomes are in the hypothesized direction but are not statistically significant at conventional levels. For issue polarization, we detect a reduction of 0.05 SD (p-value = 0.19); for candidate favorability, we recover a similar reduction of 0.03 SD (p-value 0.31).

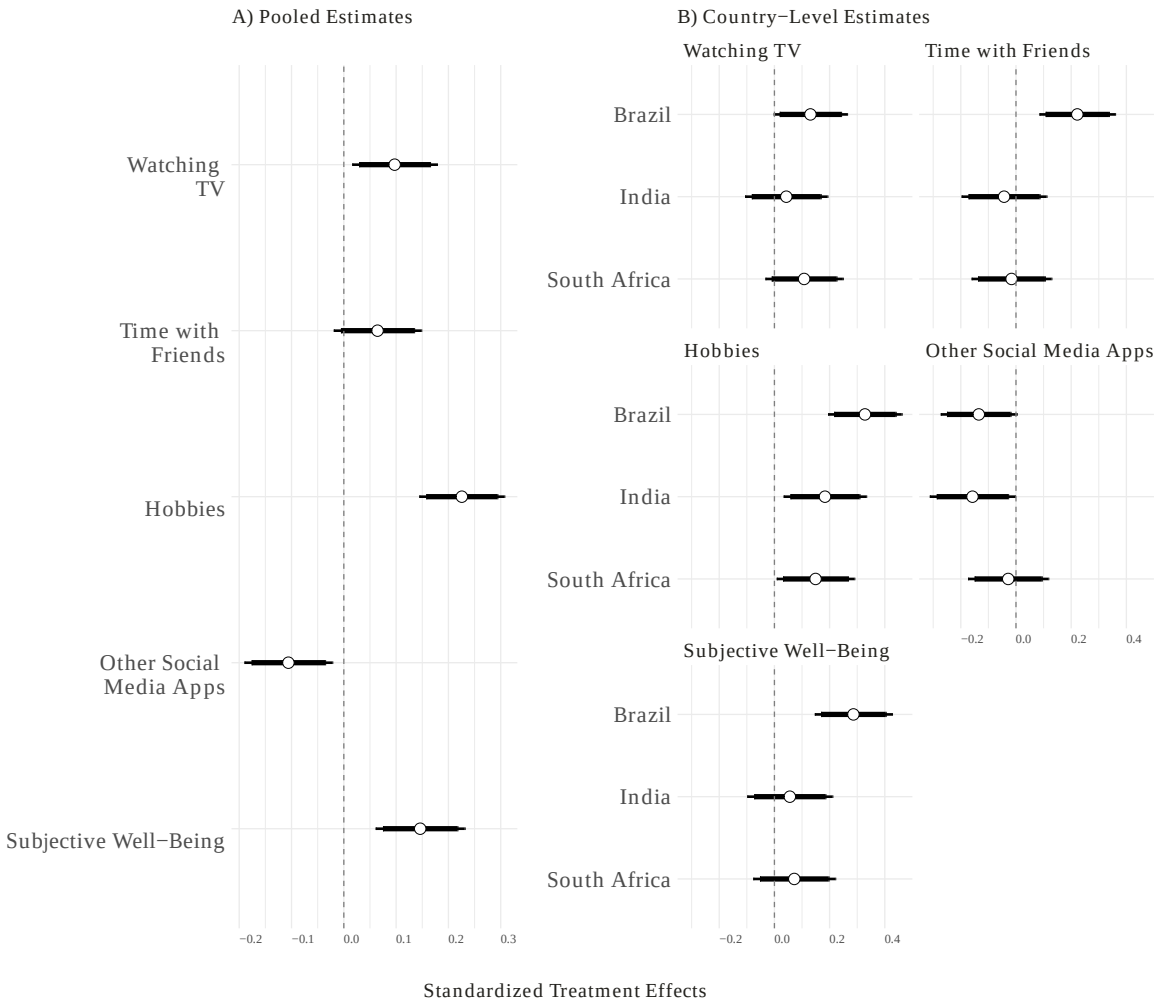
As additional research questions, we also examine downstream consequences on respondents' political interests and turnout. These outcomes, for which we did not pre-register directional hypotheses, largely mirror the political outcomes discussed thus far. In Appendix E.4, we show that our treatments failed to shift either political interest or intentions to turn out to vote. While these outcomes are exploratory, the null effects across all political measures discussed thus far suggest that simply reducing WhatsApp usage is ineffective for shifting broader political behaviors and attitudes.

[13] This measure adopts the issue agreement measure proposed in Allcott et al. (2020). In the original measure, issue polarization takes as references issue agreement for every partisan group. In our context, this is a faulty measure because several of these issues do not follow partisan cleavages. For example, in India and Brazil, only three out of the six issues exhibit a partisan cleavage among voters' issue preferences, while in South Africa, we see no issue following the partisan cleavage between the two main political coalitions in the country. Therefore, to adapt this measure to a context of fluid partisan labels, we rely on the average issue agreement in the entire control group as our reference point

[14] In India, we ask about Prime Minister Narendra Modi (BJP), Mallikarjun Kharge (Congress), Rahul Gandhi (Congress), and Yogi Adityanath (BJP). In South Africa, we ask about the leaders of four major parties running for power: Cyril Ramaphosa (ANC), who is currently president of the country, John Steenhuisen (DA), Julius Malema (EFF), and Jacob Zuma (MK). In Brazil, we ask about Lula (PT), who is currently the President, Jair Bolsonaro (PL), and the mayor of the respondent's city.

[15] In Brazil, we only consider the difference between Lula and Bolsonaro, excluding favorability towards one's mayor.

Figure 4: Non-Political Downstream Effects of Reducing WhatsApp Usage



4.4 Non-Political Downstream Effects

While reducing WhatsApp usage did not ultimately shift political outcomes, our treatments may have elicited changes at the individual level that were not political in nature. In what follows, we consider (1) substitution away from WhatsApp to different activities, and (2) changes in subjective well-being.

4.4.1 Substitution Effects

In the post-treatment survey, we included a battery of items to capture substitution effects. Primarily, we asked participants how often for the past four weeks they engaged in four different activities: (a) watching TV, (b) participating in social activities with friends, (c) spending time on offline hobbies (such as reading, crafting, sports), and (d) using other social media platforms (such as Twitter/X, Facebook, and Instagram).

Figure 4 presents the effects of reducing WhatsApp usage on the use of substitutes. Our incentives to reduce WhatsApp usage pushed individuals to substitute their time with offline activities. For instance, we detect an increase in 0.09 SD (p-value < 0.01) in watching TV, 0.06 SD (p-value $= 0.13$) in participating in social activities with friends, and 0.22 SD (p-value < 0.01) in spending time on other hobbies. Meanwhile, reducing WhatsApp usage also led participants to reducing their self-reported usage of other social media apps in the aggregated by 0.10 SD (p-value < 0.01). However, these effects are not concentrated in a specific platform. In SM figure E13, we use another battery of questions, in which we asked usage across specific social media platforms, and overall we see null effects for specific social media applications, including Telegram, Facebook, Tiktok, Twitter, and others. We see this finding as evidence that participants reduced their WhatsApp usage, and extrapolated this reduction to the overall social media time. These substitution effects are stronger in Brazil, which is consistent with the higher compliance rates we detected there, but they overall go in a similar direction in India and South Africa as well.

4.4.2 Subjective Well-Being

Participants were asked to rate how often they felt (a) anxious, (b) depressed, (c) satisfied with life, (d) happy with their appearance, and (e) isolated from family in the past few weeks on a five-point scale ranging from "Never" to "All the time." We aggregate their responses to these five items into a `Subjective Well-Being` index.[16] This constitutes the final outcome in Figure 4.

Overall, we show that reducing WhatsApp usage did improve participants' subjective well-being. Compared to those in the control condition, treated participants reported an 0.14 SD increase in subjective well-being (p-value < 0.01). However, when disaggregating the results in

[16] Negative indicators (anxiety, depression, and isolation) are reverse-coded in the index such that higher values correspond to improvements in well-being.

Figure 3(B), we note that this effect is primarily driven by participants in Brazil, while we detect no statistically significant changes in India or South Africa. This cross-country difference may be attributed to how participants in Brazil shifted their behaviors more strongly in comparison to those in India and South Africa: as we show in the previous analysis, treated participants in Brazil reported a substantial increase in offline activities and reduced social media usage overall.

4.5 Heterogeneous Treatment Effects

This section provides a brief overview of our pre-registered heterogeneous treatment effects for the primary outcomes discussed earlier. Our pre-registered moderators are digital literacy, age, overall WhatsApp usage, and WhatsApp usage for news and politics. Results are presented in the supplemental materials, section E.6. Across these moderators, the most significant conditional effects relate to overall WhatsApp usage. Figure E14 plots the conditional marginal effects for users above and below the median self-reported daily time on WhatsApp.

For participants who self-report using WhatsApp above the median user in our sample, we observe overall stronger effects on all of the informational outcomes – more substantial declines in recall of misinformation and true news headlines, as well a more substantial reduction in exposure to online hostility and low-quality political discussions. Most interestingly, we observe significant improvements in their ability to identify misinformation, suggesting enhanced discernment and indicating that, for these users, a reduction in exposure to false information online can lead to improvements in truth discernment (Pennycook, Cannon and Rand 2018). Lastly, for low-dose WhatsApp users, informational effects are more modest, except for a detectable decline in news knowledge. This effect suggests that high-dose users are likely more politically interested and obtain their news elsewhere. In contrast, low-dose users may exclusively use WhatsApp for news, resulting in a more substantial adverse effect on their knowledge. WhatsApp usage for news and politics specifically produces weak moderation effects, in a similar direction to overall usage. The conditional effects for age and digital literacy are null.

5 Conclusion

There is widespread concern that social media plays a crucial role in spreading online misinformation, creating echo chambers, spurring uncivil and toxic speech, and fomenting political polarization. Outside of Western countries, WhatsApp has been central to these concerns. Yet empirical evidence about the causal effects of WhatsApp usage is scarce, resulting in a large gap between popular accounts of WhatsApp's role in politics and academic evidence related to these concerns. Modeled after a recent wave of "deactivation" designs (Ventura et al. 2025; Asimovic et al. 2021; Allcott et al. 2020, 2024), we developed a novel field experiment geared towards reducing WhatsApp usage with two distinct treatment arms, one designed to reduce overall time spent on the app and the other to restrict the consumption of multimedia content. We deployed our experiment in three large Global South democracies—India, South Africa, and Brazil—during major elections in these countries in 2024.

Our results show that our interventions consistently reduced participants' exposure to misinformation rumors circulating in the weeks before elections. We also observe significant reductions in participants' exposure to uncivil and toxic political discourse. By incentivizing users to refrain from consuming multimedia content or limit their overall usage, we were able to decrease engagement with this type of content in the lead-up to critical elections. These results align with the conventional wisdom that social media platforms, particularly messaging apps like WhatsApp, contribute to the dissemination of false and inflammatory content (Saha et al. 2021; Tardáguila, Benevenuto and Ortellado 2018; Mello 2020). At the same time, we also establish that people rely on these platforms to some degree for receiving true information and news: compared to those who used WhatsApp as usual, treated users were less likely to recall true news headlines from the election period. We further observe that, consistent with the findings of Aslett et al. (2022) and Ventura et al. (2025), our results are strongest among those who were heavy WhatsApp users to begin with. Taken together, these findings highlight a critical trade-off: interventions that successfully reduce exposure to misinformation may simultaneously limit access to legitimate political information that citizens need for informed democratic participation. This finding is in line with studies highlighting the potential spillover effects of interventions to reduce belief in misinformation (Hoes et al. 2024).

Reductions in exposure to political information, however, did not mechanically translate to improvements in the accuracy judgments of that information. Furthermore, reducing recall of misinformation rumors, uncivil discourse, low-quality discussion, and legitimate news did not lead to meaningful or consistent changes in opinions about partisan groups, ethnic/racial groups, political issues, or political candidates. Even when our interventions successfully reduced participants' contact with highly vitriolic political content, their political attitudes remained largely unchanged. This complexity highlights the need for more nuanced theoretical models that account for selective exposure and the social embeddedness of political beliefs beyond digital platforms. Simple reductions in exposure may be insufficient to shift entrenched political attitudes when multiple reinforcing mechanisms across both online and offline environments contribute to their maintenance.

In addition to its political implications, our study also highlights the potentially positive *non-political* effects of reducing WhatsApp usage. At the end of the experiment period, participants in our treatment groups reported spending more time on their hobbies, with their friends and family, and watching television; they also exhibited notable improvements in subjective well-being and an overall reduction in social media usage beyond WhatsApp. These results underscore the potential human benefits of limiting social media, which is further strengthened when paired with our aforementioned results on reduced exposure to toxic and vitriolic content. We anticipate these findings being valuable for policymakers, practitioners, and scholars seeking to identify a balance between the need for robust information flows and the potential harms of unrestrained social media usage.

Participants' reflections on the experience provide a promising foundation when considering the scalability of our intervention. In response to an open-ended question about the experience, many treated participants described the experience as rewarding and unexpectedly beneficial even if quite challenging at times. Common themes emerged, including: the difficulty of avoiding an app as deeply embedded in individuals' lives as WhatsApp; the benefits of not seeing unnecessary or annoying content due to prioritizing their limited time on the app; and the advantage of gaining significant time to pursue other activities. Notably, 64% of respondents mentioned positive feelings and 52% mentioned negative ones, indicating that many participants candidly acknowledged the pros and cons of such a major lifestyle change. We present a comprehensive

analysis of participant perspectives in Appendix E.8.

Our research also makes a significant methodological contribution to the growing literature on characterizing the effects of social media. While much of the existing experimental work has focused on deactivating social media accounts entirely, our approach—essentially formulating a "partial deactivation" by imposing strict but reasonable limits on WhatsApp usage—constitutes a more realistic, nuanced, and ethical intervention. Indeed, this approach allows us to capture a more accurate representation of the effects of limiting social media usage without fully removing access to a tool that is central to participants' personal and professional lives.

This is also the first study (to our knowledge) that simultaneously investigates two distinct types of "partial deactivation" methods—that is, stopping multimedia consumption versus capping screen time. Our lessons from this exercise are instructive for future researchers and policymakers alike. For example, though we observed users in both treatment groups making considerable strides in limiting multimedia consumption/overall screen time, we found that, under the strictest definitions of "compliance," users in the Media arm were somewhat more successful. Indeed, a 10-minute daily limit on an app that is central to users' personal and professional lives may not always be met, but it can still dramatically reduce overall usage as individuals strive to meet it. Similarly, a complete ban on multimedia content will be flouted from time to time, but interventions constructing barriers to accessing viral multimedia content can certainly mitigate exposure to harmful content.

Within the broader conversation about social media usage and its downstream effects in the Global South (Budak et al. 2024; Badrinathan, Chauchard and Siddiqui 2024; Tucker et al. 2018), we are able to explore how our interventions function and shape outcomes in varied ways across three different countries. As we highlight in our results, we find several differences across India, Brazil, and South Africa. First, we find reductions in news recall primarily in Brazil and South Africa, and not in India. Second, in terms of reduced exposure to online incivility and low-quality political discussions, our effects are strongest in Brazil, moderate in India, and absent in South Africa. Third, even as we recovered null effects on most political outcomes across countries, we found that in India, where election rhetoric often featured religion and intergroup relations, limiting daily WhatsApp usage led to a slight reduction in identity-based (that is, religion-based) prejudice. While we offer suggestive explanations for why we saw these differences across coun-

tries based on pre-treatment and control group characteristics, future research to further consider how differences in country-specific factors might drive the relationship between social media and politics. Specifically, these variations may be in part due to differences in broader online and offline social and political dynamics that we cannot measure in our study. We therefore encourage future research to keep exploring country-level differences by replicating similar designs across diverse contexts.

Finally, our results raise important questions about the contextual dependence of social media interventions. By conducting our experiment during the last few weeks of election campaigning—which are characterized by heightened information flows and polarization—we construct "the hardest test" for detecting attitudinal changes. The effects of WhatsApp usage reduction interventions may vary substantially across different moments of the electoral cycle when information environments and users' receptiveness to new information differ. Future research should explore how similar interventions might function during non-election periods or at earlier stages of electoral cycles when political attitudes might be less crystallized and more amenable to change. Research on these contextual and temporal dynamics would help paint a more comprehensive picture of how social media messaging platforms shape information environments and political attitudes throughout the democratic process.

References

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer and Matthew Gentzkow. 2020. "The welfare effects of social media." *American Economic Review* 110(3):629–76.

Allcott, Hunt, Matthew Gentzkow, Winter Mason, Arjun Wilkins, Pablo Barberá, Taylor Brown, Juan Carlos Cisneros, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon et al. 2024. "The effects of Facebook and Instagram on the 2020 election: A deactivation experiment." *Proceedings of the National Academy of Sciences* 121(21):e2321584121.

Anspach, Nicolas M and Taylor N Carlson. 2020. "What to believe? Social media commentary and belief in misinformation." *Political Behavior* 42(3):697–718.

Arceneaux, Kevin, Martial Foucault, Kalli Giannelos, Jonathan Ladd and Can Zengin. 2023. The effects of Facebook access during the 2022 French presidential election: Can we incentivize citizens to be better informed and less polarized? Technical report Working Paper.

Arceneaux, Kevin and Martin Johnson. 2013. *Changing minds or changing channels?: Partisan news in an age of choice*. University of Chicago Press.

Aruguete, Natalia, Ernesto Calvo and Tiago Ventura. 2023. "News by popular demand: Ideological congruence, issue salience, and media reputation in news sharing." *The International Journal of Press/Politics* 28(3):558–579.

Arugute, Natalia, Ernesto Calvo and Tiago Ventura. 2022. "Network activated frames: content sharing and perceived polarization in social media." *Journal of Communication* .

Asimovic, Nejla, Jonathan Nagler and Joshua A Tucker. 2023. "Replicating the effects of Facebook deactivation in an ethnically polarized setting." *Research & Politics* 10(4):20531680231205157.

Asimovic, Nejla, Jonathan Nagler, Richard Bonneau and Joshua A Tucker. 2021. "Testing the effects of Facebook usage in an ethnically polarized setting." *Proceedings of the National Academy of Sciences* 118(25).

Aslett, Kevin, Andrew M Guess, Richard Bonneau, Jonathan Nagler and Joshua A Tucker. 2022.

"News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions." *Science advances* 8(18):eabl3844.

Avelar, Daniel. 2019. "WhatsApp fake news during Brazil election 'favoured Bolsonaro'." *The Guardian* 30:2019.

Badrinathan, Sumitra and Simon Chauchard. 2023. ""I Don't Think That's True, Bro!" Social Corrections of Misinformation in India." *The International Journal of Press/Politics* 0(0):19401612231158770.
URL: https://doi.org/10.1177/19401612231158770

Badrinathan, Sumitra, Simon Chauchard and Niloufer Siddiqui. 2024. "Misinformation and support for vigilantism: An experiment in India and Pakistan." *American Political Science Review* pp. 1–19.

Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences* 115(37):9216–9221.

Banks, Antoine, Ernesto Calvo, David Karol and Shibley Telhami. 2021. "# polarizedfeeds: Three experiments on polarization, framing, and social media." *The International Journal of Press/Politics* 26(3):609–634.

Batista Pereira, Frederico, Natália S. Bueno, Felipe Nunes and Nara Pavão. 2023. "Fake News, Fact Checking, and Partisanship: The Resilience of Rumors in the 2018 Brazilian Elections." *The Journal of Politics* 0(0):null.
URL: https://doi.org/10.1086/719419

Bessone, Pedro, Filipe R Campante, Claudio Ferraz and Pedro Souza. 2022. Social media and the behavior of politicians: Evidence from Facebook in Brazil. Technical report National Bureau of Economic Research.

Blair, Robert A, Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote and Charlene J Stain-

field. 2024. "Interventions to counter misinformation: Lessons from the Global North and applications to the Global South." *Current Opinion in Psychology* 55:101732.

Boczkowski, Pablo J, Eugenia Mitchelstein and Mora Matassi. 2018. ""News comes across when I'm in a moment of leisure": Understanding the practices of incidental news consumption on social media." *New media & society* 20(10):3523–3539.

Bor, Alexander and Michael Bang Petersen. 2022. "The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis." *American political science review* 116(1):1–18.

Bowles, Jeremy, Horacio Larreguy and Shelley Liu. 2020. "Countering misinformation via WhatsApp: Evidence from the COVID-19 pandemic in Zimbabwe." *CID Working Paper Series* .

Bowles, Jeremy, Kevin Croke, Horacio Larreguy, John Marshall and Shelley Liu. 2025. "Sustaining exposure to fact-checks: Misinformation discernment, media consumption, and its political implications." *American Political Science Review* .

Bradley, Samuel D, James R Angelini and Sungkyoung Lee. 2007. "Psychophysiological and memory effects of negative political ads: Aversive, arousing, and well remembered." *Journal of Advertising* 36(4):115–127.

Budak, Ceren, Brendan Nyhan, David M Rothschild, Emily Thorson and Duncan J Watts. 2024. "Misunderstanding the harms of online misinformation." *Nature* 630(8015):45–53.

Burgos, Pedro. 2019. "What 100,000 WhatsApp messages reveal about misinformation in Brazil." *First Draft* 27.

Chauchard, Simon and Kiran Garimella. 2022. "What Circulates on Partisan WhatsApp in India? Insights from an Unusual Dataset." *Journal of Quantitative Description: Digital Media* 2.

Cheeseman, Nic, Jonathan Fisher, Idayat Hassan and Jamie Hitchen. 2020. "Social media disruption: Nigeria's WhatsApp politics." *Journal of Democracy* 31(3):145–159.

Chen, Gina Masullo. 2017. *Online incivility and public debate: Nasty talk*. Springer.

de Freitas Melo, Philipe, Carolina Coimbra Vieira, Kiran Garimella, Pedro OS Melo and Fabrício Benevenuto. 2019. Can WhatsApp counter misinformation by limiting message forwarding? In *International conference on complex networks and their applications*. Springer pp. 372–384.

Druckman, James N, Erik Peterson and Rune Slothuus. 2013. "How elite partisan polarization affects public opinion formation." *American political science review* 107(1):57–79.

Druckman, James N, Samara Klar, Yanna Krupnikov, Matthew Levendusky and John Barry Ryan. 2021. "Affective polarization, local contexts and public opinion in America." *Nature human behaviour* 5(1):28–38.

Druckman, James N, Suji Kang, James Chu, Michael N. Stagnaro, Jan G Voelkel, Joseph S Mernyk, Sophia L Pink, Chrystal Redekopp, David G Rand and Robb Willer. 2023. "Correcting misperceptions of out-partisans decreases American legislators' support for undemocratic practices." *Proceedings of the National Academy of Sciences* 120(23):e2301836120.

Ecker, Ullrich KH, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga and Michelle A Amazeen. 2022. "The psychological drivers of misinformation belief and its resistance to correction." *Nature Reviews Psychology* 1(1):13–29.

Flaxman, Seth, Sharad Goel and Justin M Rao. 2016. "Filter bubbles, echo chambers, and online news consumption." *Public opinion quarterly* 80(S1):298–320.

Garimella, Kiran and Dean Eckles. 2020. "Images and misinformation in political groups: Evidence from WhatsApp in India." *arXiv preprint arXiv:2005.09784* .

Garimella, Kiran and Gareth Tyson. 2018. Whatapp doc? a first look at whatsapp public group data. In *Twelfth international AAAI conference on Web and Social Media*.

Ghanem, Dalia, Sarojini Hirshleifer and Karen Ortiz-Becerra. 2023. "Testing attrition bias in field experiments." *Journal of Human resources* .

Gil de Zúñiga, Homero, Alberto Ardèvol-Abreu and Andreu Casero-Ripollés. 2021. "WhatsApp political discussion, conventional participation and activism: exploring direct, indirect and generational effects." *Information, communication & society* 24(2):201–218.

Goyanes, Manuel, Alberto Ardèvol-Abreu and Homero Gil de Zúñiga. 2023. "Antecedents of news avoidance: competing effects of political interest, news overload, trust in news media, and "news finds me" perception." *Digital Journalism* 11(1):1–18.

Graham, Matthew H and Milan W Svolik. 2020. "Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States." *American Political Science Review* 114(2):392–409.

Guess, Andrew, Brendan Nyhan and Jason Reifler. 2018. "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign.".

Guess, Andrew M., Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud and Joshua A. Tucker. 2023. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science* 381(6656):398–404.
URL: https://www.science.org/doi/abs/10.1126/science.abp9364

Hall, Jeffrey A, Chong Xing, Elaina M Ross and Rebecca M Johnson. 2021. "Experimentally manipulating social media abstinence: results of a four-week diary study." *Media Psychology* 24(2):259–275.

Hanley, Sarah M, Susan E Watt and William Coventry. 2019. "Taking a break: The effect of taking a vacation from Facebook and Instagram on subjective well-being." *Plos one* 14(6):e0217743.

Haque, Md Mahfuzul, Mohammad Yousuf, Ahmed Shatil Alam, Pratyasha Saha, Syed Ishtiaque Ahmed and Naeemul Hassan. 2020. "Combating misinformation in Bangladesh: Roles and responsibilities as perceived by journalists, fact-checkers, and users." *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW2):1–32.

Hill, Seth J and Margaret E Roberts. 2023. "Acquiescence bias inflates estimates of conspiratorial beliefs and political misperceptions." *Political Analysis* 31(4):575–590.

Hoes, Emma, Brian Aitken, Jingwen Zhang, Tomasz Gackowski and Magdalena Wojcieszak. 2024. "Prominent misinformation interventions reduce misperceptions but increase scepticism." *Nature Human Behaviour* 8(8):1545–1553.

Horowitz, Donald L. 1993. "The challenge of ethnic conflict: democracy in divided societies." *Journal of democracy* 4(4):18–38.

Iyengar, Shanto, Gaurav Sood and Yphtach Lelkes. 2012. "Affect, not ideologya social identity perspective on polarization." *Public opinion quarterly* 76(3):405–431.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra and Sean J Westwood. 2019. "The origins and consequences of affective polarization in the United States." *Annual review of political science* 22(1):129–146.

Jenke, Libby. 2024. "Affective polarization and misinformation belief." *Political Behavior* 46(2):825–884.

Jones, Marc Owen. 2022. *Digital authoritarianism in the Middle East: Deception, disinformation and social media*. Oxford University Press.

Kingzette, Jon, James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky and John Barry Ryan. 2021. "How affective polarization undermines support for democratic norms." *Public Opinion Quarterly* 85(2):663–677.

Kross, Ethan, Philippe Verduyn, Gal Sheppes, Cory K Costello, John Jonides and Oscar Ybarra. 2021. "Social media and well-being: Pitfalls, progress, and next steps." *Trends in cognitive sciences* 25(1):55–66.

Lelkes, Yphtach, Gaurav Sood and Shanto Iyengar. 2017. "The hostile audience: The effect of access to broadband internet on partisan affect." *American Journal of Political Science* 61(1):5–20.

Levendusky, Matthew S. 2013. "Why do partisan media polarize viewers?" *American journal of political science* 57(3):611–623.

Levy, Ro'ee. 2021. "Social media, news consumption, and polarization: Evidence from a field experiment." *American economic review* 111(3):831–870.

Machado, Caio, Beatriz Kira, Vidya Narayanan, Bence Kollanyi and Philip Howard. 2019. A Study of Misinformation in WhatsApp groups with a focus on the Brazilian Presidential Elections. In *Companion proceedings of the 2019 World Wide Web conference*. pp. 1013–1019.

Martel, Cameron, Gordon Pennycook and David G Rand. 2020. "Reliance on emotion promotes belief in fake news." *Cognitive research: principles and implications* 5:1–20.

Masip, Pere, Jaume Suau, Carles Ruiz-Caballero, Pablo Capilla and Klaus Zilles. 2021. "News engagement on closed platforms. Human factors and technological affordances influencing exposure to news on WhatsApp." *Digital Journalism* 9(8):1062–1084.

Matthes, Jörg, Kathrin Karsay, Desirée Schmuck and Anja Stevic. 2020. ""Too much to handle": Impact of mobile social networking sites on information overload, depressive symptoms, and well-being." *Computers in Human Behavior* 105:106217.

Mello, Patrícia Campos. 2020. *A máquina do ódio: notas de uma repórter sobre fake news e violência digital*. Companhia das Letras.

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

Newman, Nic, Richard Fletcher, Craig T Robertson, A Ross Arguedas and Rasmus Kleis Nielsen. 2024. *Reuters Institute digital news report 2024*. Reuters Institute for the study of Journalism.

Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y. Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Andrew M. Guess, Edward Kennedy, Young Mie Kim, David Lazer, Neil Malhotra, Devra Moehler, Jennifer Pan, Daniel Robert Thomas, Rebekah Tromble, Carlos Velasco Rivera, Arjun Wilkins, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud and Joshua A. Tucker. 2023. "Like-minded sources on Facebook are prevalent but not polarizing." *Nature* .

URL: https://doi.org/10.1038/s41586-023-06297-w

Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann and Michael Bang Petersen. 2021. "Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter." *American Political Science Review* pp. 1–17.

Pennycook, Gordon, Tyrone D Cannon and David G Rand. 2018. "Prior exposure increases perceived accuracy of fake news." *Journal of experimental psychology: general* 147(12):1865.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles and David G Rand. 2021. "Shifting attention to accuracy can reduce misinformation online." *Nature* 592(7855):590–595.

Persily, Nathaniel and Joshua A Tucker. 2020. *Social media and democracy: The state of the field, prospects for reform*. Cambridge University Press.

Piazza, James A. 2023. "Political polarization and political violence." *Security Studies* 32(3):476–504.

Poushter, Jacob. 2024. "WhatsApp and Facebook dominate the social media landscape in middle-income nations." *Pew Research Center* .

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

Rathje, Steve, Jay J Van Bavel and Sander Van Der Linden. 2021. "Out-group animosity drives engagement on social media." *Proceedings of the national academy of sciences* 118(26):e2024292118.

Recuero, Raquel, Felipe Soares and Otávio Vinhas. 2021. "Discursive strategies for disinformation on WhatsApp and Twitter during the 2018 Brazilian presidential election." *First Monday* .

Resende, Gustavo, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida and Fabrício Benevenuto. 2019. (Mis) information dissemination in WhatsApp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference*. pp. 818–828.

Resende, Gustavo, Philipe Melo, Julio C.S. Reis, Marisa Vasconcelos, Jussara M Almeida and Fabrício Benevenuto. 2019. Analyzing textual (mis) information shared in WhatsApp groups. In *Proceedings of the 10th ACM conference on web science*. pp. 225–234.

Rossini, Patrícia. 2022. "Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk." *Communication Research* 49(3):399–425.

Rossini, Patricia, Érica Anita Baptista, Vanessa Veiga de Oliveira and Jennifer Stromer-Galley. 2021. "Digital media landscape in Brazil: Political (Mis) information and participation on Facebook and WhatsApp." *Journal of Quantitative Description: Digital Media* 1.

Rossini, Patrícia, Jennifer Stromer-Galley, Erica Anita Baptista and Vanessa Veiga de Oliveira. 2021. "Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections." *New Media & Society* 23(8):2430–2451.

Rowe, Ian. 2015. "Civility 2.0: A comparative analysis of incivility in online political discussion." *Information, communication & society* 18(2):121–138.

Saha, Punyajoy, Binny Mathew, Kiran Garimella and Animesh Mukherjee. 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.

Settle, Jaime E. 2018. *Frenemies: How social media polarizes America*. Cambridge University Press.

Stroud, Natalie Jomini. 2011. *Niche news: The politics of news choice*. Oxford University Press.

Sunstein, Cass R. 2018. # *Republic*. Princeton university press.

Svolik, Milan W. 2019. "Polarization versus democracy." *Journal of democracy* 30(3):20–32.

Tardáguila, Cristina, Fabricio Benevenuto and Pablo Ortellado. 2018. "Fake news is poisoning Brazilian politics. WhatsApp can stop it." *The New York Times* 17(10).

Tokita, Christopher K, Andrew M Guess and Corina E Tarnita. 2021. "Polarized information ecosystems can reorganize social networks via information cascades." *Proceedings of the National Academy of Sciences* 118(50):e2102147118.

Tromholt, Morten. 2016. "The Facebook experiment: Quitting Facebook leads to higher levels of well-being." *Cyberpsychology, behavior, and social networking* 19(11):661–666.

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social media, political polarization, and political disinformation: A review of the scientific literature." *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* .

Tucker, Joshua A, Yannis Theocharis, Margaret E Roberts and Pablo Barberá. 2017. "From liberation to turmoil: Social media and democracy." *J. Democracy* 28:46.

Twenge, Jean M and W Keith Campbell. 2018. "Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study." *Preventive medicine reports* 12:271–283.

Valenzuela, Sebastián, Ingrid Bachmann and Matías Bargsted. 2021. "The personal is the political? What do Whatsapp users share and how it matters for news knowledge, polarization and participation in Chile." *Digital journalism* 9(2):155–175.

Vanman, Eric J, Rosemary Baker and Stephanie J Tobin. 2018. "The burden of online friends: The effects of giving up Facebook on stress and well-being." *The Journal of social psychology* 158(4):496–508.

Velez, Yamil Ricardo and Patrick Liu. 2024. "Confronting core issues: A critical assessment of attitude polarization using tailored experiments." *American Political Science Review* pp. 1–18.

Ventura, Tiago, Rajeshwari Majumdar, Jonathan Nagler and Joshua A. Tucker. 2025. "Misinformation Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil." *The Journal of Politics* .

Voelkel, Jan G, Michael Stagnaro, James Chu, Sophia Pink, Joseph Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi Allen et al. 2023. "Megastudy identifying effective interventions to strengthen Americans' democratic attitudes.".

Wilkinson, Steven. 2006. *Votes and violence: Electoral competition and ethnic riots in India*. Cambridge University Press.

Wirtschafter, Valerie, Frederico Batista Pereira, Natália Bueno, Nara Pavão, Felipe Nunes et al. 2024. "Detecting misinformation: Identifying false news spread by political leaders in the global south." *Journal of Quantitative Description: Digital Media* 4.

Wojcieszak, Magdalena, Bernhard Clemm von Hohenberg, Andreu Casas, Ericka Menchen-Trevino, Sjifra de Leeuw, Alexandre Gonçalves and Miriam Boon. 2022. "Null effects of news exposure: a test of the (un) desirable effects of a 'news vacation' and 'news binging'." *Humanities and Social Sciences Communications* 9(1):1–10.

Statements

Author Contributions

R.M. and T.V. are co-first authors ordered alphabetically. R.M., T.V., S.L., and J.T. designed the study. R.M., T.V., S.L., and C.T. collected the data and supervised the fieldwork. R.M., T.V., S.L., and C.T. wrote the first draft, and all authors contributed to reviewing and editing it.

Funding Statement:

This research was funded by the McCourt School's Tech & Policy Program and Project Liberty's Institute. We also gratefully acknowledge that The Center for Social Media and Politics at New York University is supported by funding from the John S. and James L. Knight Foundation, the Charles Koch Foundation, the Hewlett Foundation, Craig Newmark Philanthropies, the William and Flora Hewlett Foundation, the Siegel Family Endowment, and the Bill and Melinda Gates Foundation.

Acknowledgments

We thank Jeremy Bowles, Nejla Asimovic, and participants at the 2024 American Political Science Association Annual Meeting, 2025 Midwest Political Science Association Annual Conference, 2025 Media Effects Workshop at Columbia University, 2025 NYU CSMaP Annual Conference, Georgetown Mortara Research Seminar, FGV-EBAPE Seminar Series, and NYU WINNING Summer Workshop for helpful feedback. We also thank Maitreyi Natarajan, Anjali Ofori, Brooks Clifford, Isabella Remor, and Sunaina Kathpalia for excellent research assistance.

Competing interests

J.T. 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. J.T. 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. J.T. is currently a Senior Geopolitical Risk Advisor at Kroll. The remaining authors declare no conflicts of interest.

Reducing Social Media Usage During Elections: Evidence from a WhatsApp Multi-Country Deactivation Experiment

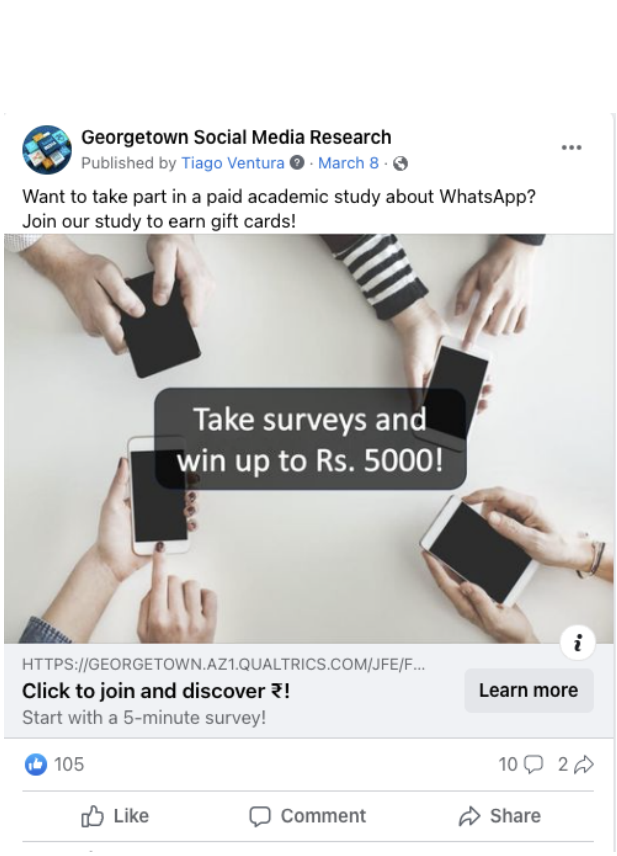
Supplemental Materials (SM)

A Experiment materials

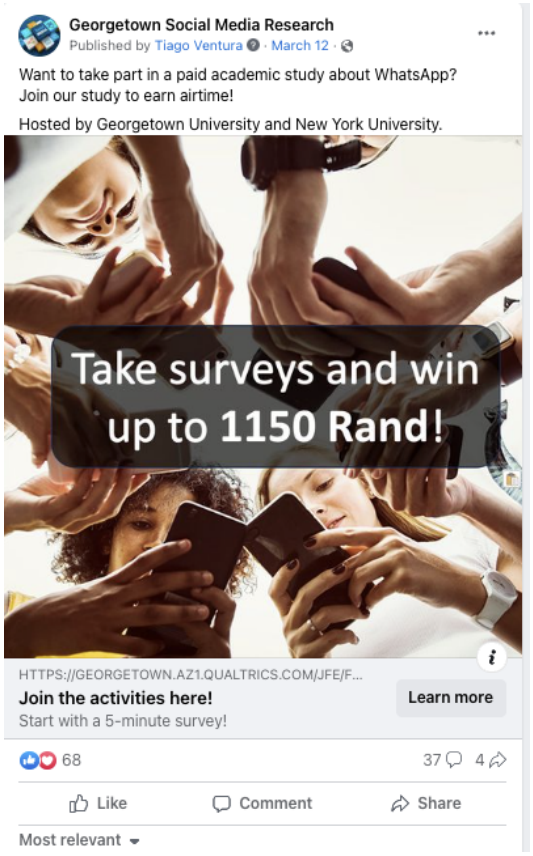
A.1 Recruitment materials

For the Facebook Ads recruitment, we used simple text and images to invite participants to a paid academic study about WhatsApp. We used multiple small variations of the text and images to increase recruitment. We present one example as respondents saw on their Facebook timelines:

Figure A5: Facebook Advertisement Used for Recruitment



a) India Ad



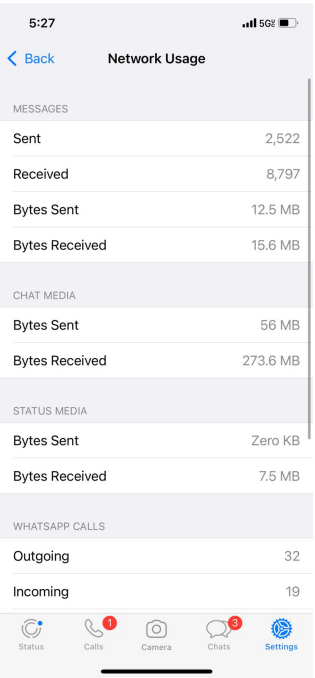
b) South Africa Ad

A.2 Usage screenshots

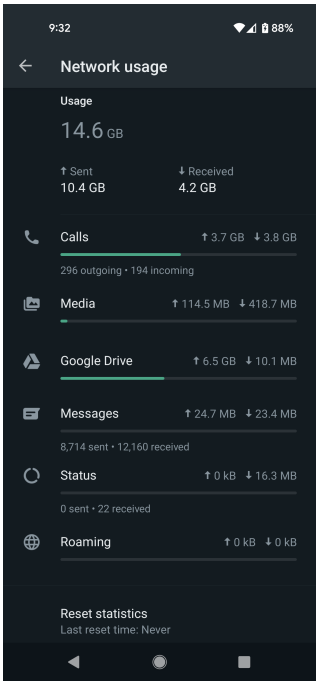
Figure A6: WhatsApp Usage Screenshot

A) iPhone Screenshots

B) Android Screenshots



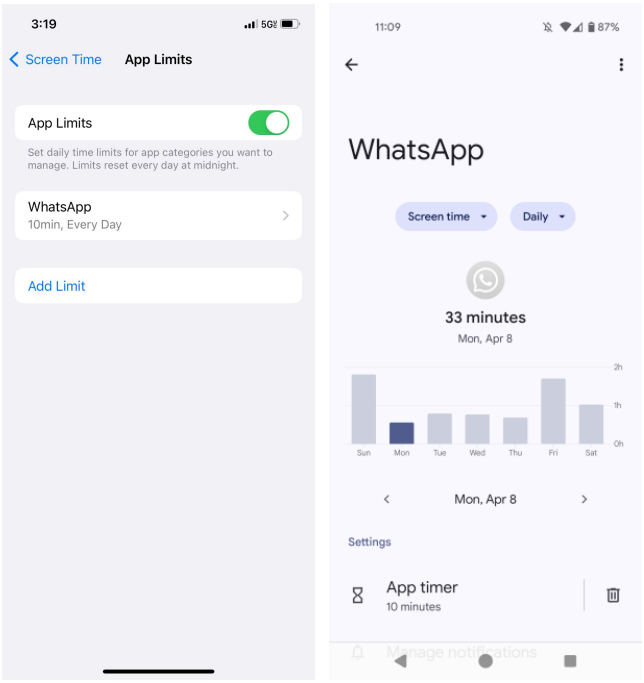
C) iPhone Screenshots



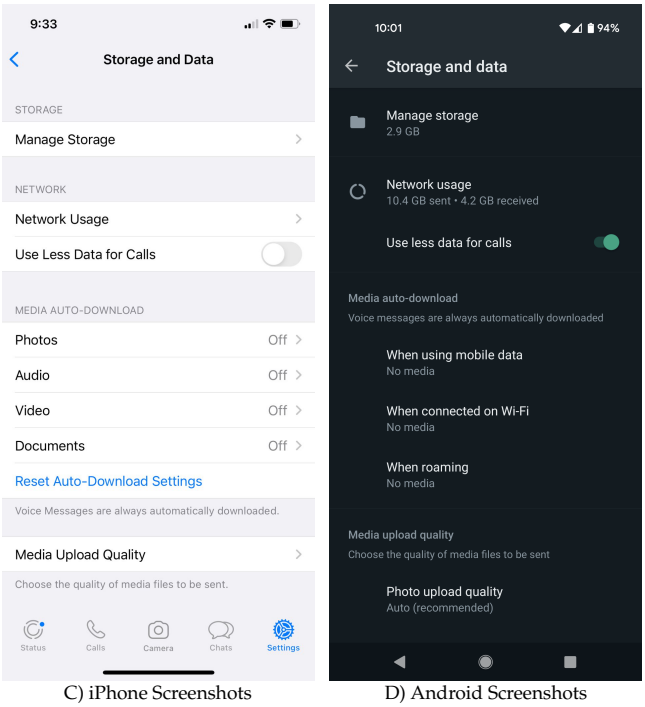
D) Android Screenshots

Note: The upper row presents examples of the screenshots of usage information requested from participants assigned to **Time**. The bottom row presents examples of the screenshots of multimedia consumption requested from participants assigned to **Multimedia**.

Figure A7: WhatsApp Treatment Screenshot



A) iPhone Screenshots B) Android Screenshots



C) iPhone Screenshots D) Android Screenshots

Note: The upper row presents examples of the screenshots requested from participants assigned to **Time**. The bottom row presents examples of the screenshots requested from participants assigned to **Multimedia**.

A.3 Incentives

Participants were compensated for every task that they completed. We present below the values in USD, but participants were paid according to the local currency. Given the differential effort required across treatment and control, we offered an additional bonus for treated participants. We inform participants that this bonus payment is conditional on their compliance with the study, which we would measure by looking at their submitted screenshots. The incentives are as follows.

- Baseline survey: 3 dollars
- Compliance task: 1 dollar for each screenshot (4 dollars total)
- Final survey: 8 dollars
- Bonus payment for treated participants who successfully comply: 45 dollars
- Lottery prize: 100 dollars

Thus, those in the control conditions can earn up to 15 dollars; those in the treatment conditions, 60 dollars. All participants are entered into a lottery for one of three 100-dollar gift cards at the end of the study.

B Outcome measures

Tables B2 and B3 describe the outcomes included in our primary analysis. Table B4 presents the headlines shown to participants in the three countries. All headlines are presented in English as seen by participants in India and South Africa. Participants in Brazil saw them in Portuguese, but we present their translated versions here.

Table B2: Pre-Registered Outcomes and Measurement Choices:Information Outcomes

Variable Name	Definition	Measurement
Misinformation Recall	Measures the self-reported recall of **Misinformation Rumors** using a headline task	Sum of the misinformation rumors recalled.
News Recall	Measures the self-reported recall of *True News* using a headline task	Sum of the true news recalled.
Misinformation Accuracy	Measures the ability to discern false information using a headline task	Sum of the **Misinformation Rumors** classified as false.
News Knowledge Accuracy	Measures the ability to discern true news information using a headline task	Sum of the news headlines (*True News* and *Placebo False News*) correctly classified as accurate or not accurate, respectively.
Online Toxicity	Composite social media incivility score	Sum of the standardized z-scores for online toxicity and online incivility measures
Quality of Political Discussions	Composite score for five-point scale of self-reported experiences of the quality of political discussion for the past four weeks	Sum of the standardized z-scores for four items: political anger, political incivility, information overload

Table B3: Pre-Registered Outcomes and Measurement Choices: Political Outcomes

Variable Name	Definition	Measurement
Polarization Index	Composite polarization measure	Sum of the z-scores for three polarization outcomes (affective polarization, social polarization, traits polarization)
Affective Polarization	Measures affective polarization towards the two major political parties/coalition in each country	Absolute value of the difference between the feeling scales for each political party/coalition
Social Polarization	Measures social polarization towards the two major political parties/coalition in each country	Absolute value of the difference between the number of social activities willing to engage with voters from each political party/coalition
Traits Polarization	Measures positive and negative traits assigned to voters from main political parties/coalition in each country	Difference between number of negative and positive traits assigned to outgroup party/coalition
Identity-based prejudice Index	Composite index measuring levels of identity-based prejudice	Sum of the z-scores for three identity polarization measures index (Affective prejudice, Social prejudice, Traits prejudice)
Affective Identity-Based Prejudice	Measures affective outgroup prejudice between participants' ingroup and outgroup in each country	Absolute value of the difference between the feeling scales for in/outgroup
Social Identity-Based Prejudice	Measures social ethnic prejudice between participants' ingroup and outgroup in each country	Absolute value of the difference between the number of social activities willing to engage with ingroup and outgroup
Traits Identity-Based Prejudice	Measures positive and negative traits assigned participants' outgroup in each country	Difference between number of negative and positive traits assigned to outgroup
Candidate Favorability	Measure overall positive feeling towards the participants' preferred candidate	Absolute difference between participants' main political ingroup candidate and the candidates from the alternative parties
Issue Polarization	Measures extremity compared to median voter with six issue opinions questions	Standardized absolute differences between participants' agreement (on each issue) to the mean issue agreement among the entire control group

B.1 Headlines

Table B4: Misinformation Rumors and True news to measure recall and belief accuracy. The order of the items was fully randomized.

Headlines	Category
Brazil	
At a rally on September 7, Bolsonaro got angry at Pablo Marçal's sound car and called the candidate a lazy person	Misinformation Rumors
In a video, journalist Sandra Annenberg announced the resgata brasil program, where Brazilians can win up to 7,000 reais in compensation due to a leakage of credit card data	Misinformation Rumors
In a recent decision, Finance Minister Fernando Haddad announced 0nd of unemployment insurance for 2025	Misinformation Rumors
Commenting on the elections in Venezuela, Lula criticizes the electronic ballot box, and defends the use of printed ballots by Venezuelan Authorities	Misinformation Rumors
For drug trafficking and corruption with public money, court orders arrest of singer Gustavo Lima and influencer Deolane Bezerra	Placebo News
After negotiations in New York, Israel, Lebanon and Palestine reach peace agreement and cease attacks in the region.	Placebo News
Government announces that 600 betting sites will be banned from Brazil in October	True News
In New York, President Lula receives award from billionaire Bill Gates for policies to combat hunger	True News
Alexandre de Moraes last week denied Twitter/X's return to Brazil despite the company complying with some of the court's requests	True News
Filhes desse solo: National Anthem is sung in neutral language at Guilherme Boulos' rally with Lula.	True News
India	
Muslim women were caught doing fake voting under their burqas during the Lok Sabha elections	Misinformation Rumors
Congress promises that the money of regular Indian citizens will be collected and distributed to poor Muslims	Misinformation Rumors
No new Public Sector Enterprises have been incorporated under the Modi Govt	Misinformation Rumors
BJP is using Army personnel to influence citizens in election booths to vote for the BJP	Misinformation Rumors
Rahul Gandhi praises PM Modi for his quick response to Pune Porsche accident	Placebo News
New study finds population rise is related to religion, highest increase in fertility rate among Muslims	Placebo News
Modi accuses Opposition of using "vote jihad" to win elections	True News
Arvind Kejriwal will have to surrender and go back to jail on June 2	True News
Congress leaders call PM Modi a dictator	True News
Election-time seizures of cash, drugs, liquor to cross all-time high mark in 2024	True News
South Africa	
If you are registered the vote but don't vote on elections day, then the ANC will receive your vote automatically.	Misinformation Rumors
Mozambican migrants are being imported into South Africa to vote for the ANC.	Misinformation Rumors
You can get a South African ID for R4,500 quickly by applying through WhatsApp.	Misinformation Rumors
South Africa's biggest trade union, NUMSA, has asked its members to vote for the MK Party.	Misinformation Rumors
Political leaders have called DA's burning of SA flag a 'heroic act'.	Placebo News
Jacob Zuma agreed to step down from the race after being locked out of South Africa elections.	Placebo News
Parliament gives Ramaphosa a blank cheque to set donation limits	True News
DA accuses Rise Mzansi of fueling racial tensions in Western Cape.	True News
The South Africa Medical Association is planning to mount legal challenge against the new National Health Insurance law.	True News
Jabulani Khumalo takes fight to remove Jacob Zuma as MK Party leader to Electoral Court.	True News

C Sample Characteristics and Attrition Analysis

We invited 6,261 participants to take part in the WhatsApp Reduction Intervention (2,067 in BR, 1,310 in IN, 2,884 in SA). Out of the individuals invited, 2,425 participants were successfully enrolled in the experiment (926 in Brazil, 679 in India, and 820 in South Africa). Then, out of the participants who started the interventions, 2,220 completed the post-treatment survey (825 in Brazil, 653 in India, and 742 in South Africa). We divide this section into three parts. First, we present the baseline characteristics of our enrolled sample with benchmarks from other population surveys in Brazil, India, and South Africa. Second, we compare baseline characteristics between the participants we invited and those who enrolled in the experiment. We called this analysis *selection among the enrolled*. Understanding this pattern helps us calibrate the external validity of our findings. Lastly, we conduct a series of attrition analyses comparing baseline differences between participants who successfully enrolled in the study, and participants who attrited from the study between enrollment and post-treatment survey. We present a set of formal tests of attrition bias as suggested in Ghanem, Hirshleifer and Ortiz-Becerra (2023) comparing these sample

C.1 Selection Among the Enrolled

In this section, we report differences in the sample characteristics between participants invited to those who successfully enrolled in the study. Invited participants are those who pass all the eligibility criteria discussed in the main manuscript, and we sent at least one email or WhatsApp message inviting them to return to the study. All eligible participants in India and South Africa were invited, while in Brazil, we achieved our sample size before inviting all eligible. We use the following variables (all converted to numerical scales) to understand differences between these groups: age, gender, education, time spent on WhatsApp, use of only mobile WhatsApp, WhatsApp usage for chatting with friends, WhatsApp usage for news consumption, WhatsApp usage for work-related tasks, and exposure to multimedia about politics on WhatsApp.

Table C5 presents the results. For each of the eight variables, we provide pairwise t-tests and omnibus F-statistics for the entire model to examine the joint orthogonality for all eight variables. We detect statistically significant differences for most of the variables. Overall, participants that enrolled in the experiments were younger, more educated, used to spend more time on WhatsApp,

Table C5: Baseline Characteristics Between Invited and Enrolled Participants

	Dropout (N=3836)		Enrolled (N=2425)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
Age	2.62	1.22	2.35	1.05	-0.27	<0.01
Gender	0.42	0.50	0.43	0.50	0.01	0.60
Education	4.47	0.92	4.69	0.85	0.22	<0.01
WP: Daily time	3.49	1.48	3.61	1.40	0.13	<0.01
WP: Mobile Only	0.86	0.35	0.71	0.46	-0.15	<0.01
WP: Use for Work	5.12	1.17	5.09	1.05	-0.02	0.40
WP: Use for News	4.99	1.23	4.78	1.27	-0.21	<0.01
WP: Use for Chat with Friends	5.51	0.88	5.44	0.87	-0.06	<0.01
WP:Frequency Images about Politics	3.92	1.54	4.14	1.47	0.21	<0.01
Omnibus Test: F-Statistics = 49.8 p-values <0.01						

Table C6: Baseline Characteristics Between Treatment and Control Participants Enrolled

	Control (N=1198)		Treatment (N=1227)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
Age	2.36	1.08	2.34	1.03	-0.02	0.72
Gender	0.42	0.50	0.44	0.51	0.02	0.32
Education	4.71	0.85	4.67	0.84	-0.04	0.23
WP: Daily time	3.63	1.40	3.60	1.41	-0.02	0.67
WP: Mobile Only	0.70	0.46	0.71	0.45	0.02	0.36
WP: Use for Work	5.11	1.04	5.07	1.06	-0.04	0.35
WP: Use for News	4.79	1.27	4.77	1.28	-0.01	0.78
WP: Use for Chat with Friends	5.44	0.88	5.45	0.85	0.01	0.89
WP:Frequency Images about Politics	4.14	1.49	4.13	1.46	-0.01	0.90
Omnibus Test: F-Statistics = 0.49, p-values = 0.893						

are less likely to be mobile-only WhatsApp users, less likely to use WhatsApp for news and to talk with friends, but more likely to receive political content through multimedia on WhatsApp. These differences contribute to high and statistically significant F-statistics. We use the same set of covariates to examine differences between the treatment and control enrolled participants. Table C6 shows the results, and we do not find statistically significant differences in any of the eight variables, as well as in the omnibus F-test. Therefore, while we do see evidence of selection in the experiment, particularly with a likely more digitally savvy group joining the experiment, we do not find any evidence of this issue affecting the balance between treatment and control.

C.2 Attrition Analysis

In this section, we report findings related to attrition rates within the intervention. To address the potential for attrition bias, we apply statistical tests introduced by Ghanem, Hirshleifer and Ortiz-

Becerra (2023). Rather than focusing solely on baseline comparisons between the treatment and control groups, this approach offers a more formal framework for evaluating attrition bias in field experiments by leveraging baseline outcome data. Specifically, Ghanem, Hirshleifer and Ortiz-Becerra (2023) demonstrate that the key identifying assumption for estimating local treatment effects can be assessed by testing two equality conditions: one concerning the baseline outcome distributions of the treatment and control groups, and another concerning the same distributions among those who attrited from each group. We use the following baseline covariates and proxies for the outcomes: age, gender, education, income, self-reported WhatsApp usage per day, news consumption, news consumption on social media, self-reported exposure to false information, partisan affective polarization, and ethnic prejudice. The last two are calculated using the absolute difference between the two main political and ethnic/racial groups across the three countries, except for Brazil which we did not measure the ethnic prejudice at baseline.

Table C7 examines the presence of attrition bias using this recommended approach. Column 1 reports the attrition rate for control, and Column 2 reports the differential attrition rate between treatment and control, with the corresponding p-value testing for difference in attrition between the groups (*differential attrition*). We find no evidence of differential attrition in the full sample. Columns 3-6 present the mean baseline outcome for treatment respondents (TR), control respondents (CR), treatment attriters (TA), and control attriters (CA), respectively. Column 7 reports the p-value of the hypothesis test with two equality restrictions (*selective attrition*. We cannot reject the joint null hypothesis of no differences in the mean outcome baseline values across treatment and control respondents, as well as treatment and control attritors. This test indicates that selective attrition is not a threat to the experiment's internal validity.

Tables C8, C9, C10 present the same results split for each country. In Brazil, only one out of the ten variables do not pass the selective attrition test. In India, no evidence of differential attrition or selective attrition has been detected. In South Africa, no evidence of selective attrition is detected, but we find a small variation in differential attrition. Since most of the pooled results are not driven by the South African sample, we do not consider this a threat to the internal validity of our design.

Table C7: Sample Internal Validity Test for Primary Outcomes and Baseline Variables

	Attrition Bias		Mean baseline outcome by group				Test of internal validity
Outcome	C	Differential	TR	CR	TA	CA	pvalue
Age	0.91	0.002 (p-value = 0.8577)	2.33	2.35	2.43	2.44	0.94
Gender	0.91	0.002 (p-value = 0.8577)	0.44	0.42	0.47	0.44	0.60
Education	0.91	0.002 (p-value = 0.8577)	4.68	4.72	4.51	4.58	0.47
Income	0.91	0.002 (p-value = 0.8577)	4.54	4.64	3.88	3.72	0.36
WP:Daily time	0.91	0.002 (p-value = 0.8577)	3.61	3.65	3.49	3.37	0.68
News Consumption: General	0.91	0.002 (p-value = 0.8577)	3.57	3.61	2.81	2.82	0.77
News Consumption: Social Media Apps	0.91	0.002 (p-value = 0.8577)	0.93	0.92	0.90	0.87	0.57
False News Exposure	0.91	0.002 (p-value = 0.8577)	2.49	2.44	2.51	2.50	0.57
Affective Polarization	0.91	0.002 (p-value = 0.8577)	-0.03	0.01	0.23	0.01	0.16
Ethnic Prejudice	0.91	0.002 (p-value = 0.8577)	0.11	0.16	0.07	0.18	0.61

Note: Column 1 reports the attrition rate for control, and Column 2 reports the differential attrition rate between treatment and control, with the corresponding p-value testing for difference in attrition between the groups (*differential attrition*). Columns 3-6 present the mean baseline outcome for treatment respondents (TR), control respondents (CR), treatment attriters (TA), and control attriters (CA), respectively. Column 7 reports the p-value of the hypothesis test with two equality restrictions (*selective attrition*

Table C8: Brazil Sample Internal Validity Test for Primary Outcomes and Baseline Variables

	Attrition Bias		Mean baseline outcome by group				Test of internal validity
Outcome	C	Differential	TR	CR	TA	CA	pvalue
Age	0.87	0.0317 (p-value = 0.1236)	2.50	2.51	2.63	2.56	0.95
Gender	0.87	0.0317 (p-value = 0.1236)	0.35	0.35	0.44	0.41	0.94
Education	0.87	0.0317 (p-value = 0.1236)	4.49	4.52	4.30	4.46	0.40
Income	0.87	0.0317 (p-value = 0.1236)	5.00	5.13	4.30	3.98	0.38
WP:Daily time	0.87	0.0317 (p-value = 0.1236)	3.79	3.82	3.51	3.41	0.89
News Consumption: General	0.87	0.0317 (p-value = 0.1236)	3.11	3.12	2.51	2.69	0.78
News Consumption: Social Media Apps	0.87	0.0317 (p-value = 0.1236)	0.91	0.87	0.88	0.81	0.11
False News Exposure	0.87	0.0317 (p-value = 0.1236)	2.42	2.36	2.21	2.39	0.60
Affective Polarization	0.87	0.0317 (p-value = 0.1236)	0.19	0.22	0.71	0.14	0.03

Note: Column 1 reports the attrition rate for control, and Column 2 reports the differential attrition rate between treatment and control, with the corresponding p-value testing for difference in attrition between the groups (*differential attrition*). Columns 3-6 present the mean baseline outcome for treatment respondents (TR), control respondents (CR), treatment attritters (TA), and control attritters (CA), respectively. Column 7 reports the p-value of the hypothesis test with two equality restrictions (*selective attrition*)

Table C9: India Sample Internal Validity Test for Primary Outcomes and Baseline Variables

	Attrition Bias		Mean baseline outcome by group				Test of internal validity
Outcome	C	Differential	TR	CR	TA	CA	pvalue
Age	0.95	0.0187 (p-value = 0.205)	2.22	2.18	2.10	2.56	0.51
Gender	0.95	0.0187 (p-value = 0.205)	0.69	0.67	0.60	0.50	0.75
Education	0.95	0.0187 (p-value = 0.205)	4.98	4.97	4.30	4.94	0.28
Income	0.95	0.0187 (p-value = 0.205)	4.58	4.49	4.80	4.12	0.31
WP:Daily time	0.95	0.0187 (p-value = 0.205)	3.11	3.15	3.10	2.81	0.81
News Consumption: General	0.95	0.0187 (p-value = 0.205)	4.07	3.99	3.20	2.81	0.57
News Consumption: Social Media Apps	0.95	0.0187 (p-value = 0.205)	0.94	0.96	0.90	0.94	0.39
False News Exposure	0.95	0.0187 (p-value = 0.205)	2.54	2.38	3.00	2.56	0.15
Affective Polarization	0.95	0.0187 (p-value = 0.205)	-0.10	0.05	-0.17	-0.07	0.13

Note: Column 1 reports the attrition rate for control, and Column 2 reports the differential attrition rate between treatment and control, with the corresponding p-value testing for difference in attrition between the groups (*differential attrition*). Columns 3-6 present the mean baseline outcome for treatment respondents (TR), control respondents (CR), treatment attriters (TA), and control attriters (CA), respectively. Column 7 reports the p-value of the hypothesis test with two equality restrictions (*selective attrition*)

Table C10: South Africa Sample Internal Validity Test for Primary Outcomes and Baseline Variables

	Attrition Bias		Mean baseline outcome by group				Test of internal validity
Outcome	C	Differential	TR	CR	TA	CA	pvalue
Age	0.93	-0.0463 (p-value = 0.0239)	2.25	2.32	2.32	2.11	0.42
Gender	0.93	-0.0463 (p-value = 0.0239)	0.30	0.27	0.46	0.46	0.63
Education	0.93	-0.0463 (p-value = 0.0239)	4.64	4.73	4.74	4.64	0.25
Income	0.93	-0.0463 (p-value = 0.0239)	4.00	4.22	3.34	2.93	0.26
WP:Daily time	0.93	-0.0463 (p-value = 0.0239)	3.85	3.89	3.54	3.61	0.92
News Consumption: General	0.93	-0.0463 (p-value = 0.0239)	3.62	3.83	2.98	3.07	0.14
News Consumption: Social Media Apps	0.93	-0.0463 (p-value = 0.0239)	0.94	0.94	0.92	0.96	0.73
False News Exposure	0.93	-0.0463 (p-value = 0.0239)	2.54	2.58	2.68	2.68	0.90
Affective Polarization	0.93	-0.0463 (p-value = 0.0239)	-0.23	-0.26	-0.10	-0.21	0.73

Note: Column 1 reports the attrition rate for control, and Column 2 reports the differential attrition rate between treatment and control, with the corresponding p-value testing for difference in attrition between the groups (*differential attrition*). Columns 3-6 present the mean baseline outcome for treatment respondents (TR), control respondents (CR), treatment attriters (TA), and control attriters (CA), respectively. Column 7 reports the p-value of the hypothesis test with two equality restrictions (*selective attrition*

D Robustness Models

D.1 Unadjusted Intention-to-Treat Effects

Table D11: Treatment Effects: Intention-to-Treat Unadjusted Models

Outcomes	Pooled Effects	Brazil	India	South Africa
Information Outcomes				
Misinformation Recall	-0.142 (0.039) ⋆⋆⋆	-0.14 (0.065) ⋆	-0.183 (0.073) ⋆	-0.109 (0.068)
News Recall	-0.196 (0.041) ⋆⋆⋆	-0.295 (0.068) ⋆⋆⋆	-0.042 (0.076)	-0.221 (0.072) ⋆⋆
Misinformation Accuracy	0.036 (0.041)	0.046 (0.068)	0.081 (0.076)	-0.016 (0.072)
News Accuracy	-0.029 (0.042)	-0.136 (0.068) ⋆	0.011 (0.077)	0.054 (0.072)
Online Toxicity	-0.094 (0.042) ⋆	-0.222 (0.069) ⋆⋆	-0.055 (0.078)	0.016 (0.073)
Low-Quality Political Discussions	-0.081 (0.043) ○	-0.117 (0.07) ○	-0.09 (0.079)	-0.032 (0.074)
Attitudinal Outcomes				
Partisan Polarization	-0.032 (0.043)	-0.072 (0.07)	-0.116 (0.078)	0.087 (0.074)
Identity-based Prejudice	-0.035 (0.054)	NA	-0.124 (0.079)	0.044 (0.075)
Issue Polarization	-0.065 (0.041)	-0.11 (0.068)	-0.004 (0.076)	-0.07 (0.072)
Candidate Favorability	-0.06 (0.04)	-0.103 (0.066)	-0.06 (0.073)	-0.009 (0.071)
Additional Research Questions				
Watching TV	0.089 (0.042) ⋆	0.124 (0.068) ○	0.034 (0.077)	0.098 (0.072)
Time with Friends	0.06 (0.043)	0.217 (0.071) ⋆⋆	-0.049 (0.079)	-0.02 (0.074)
Hobbies	0.22 (0.042) ⋆⋆⋆	0.323 (0.068) ⋆⋆⋆	0.178 (0.077) ⋆	0.141 (0.072) ○
Other Social Media Apps	-0.108 (0.043) ⋆	-0.136 (0.07) ○	-0.166 (0.079) ⋆	-0.026 (0.074)
Subjective Well-Being	0.146 (0.044) ⋆⋆⋆	0.283 (0.072) ⋆⋆⋆	0.058 (0.079)	0.072 (0.076)

Note: Standard errors in parentheses. All models use multilevel estimations with random intercepts by country. ○ $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

E Additional Results

E.1 Unpooled Media and Time Treatment Effects

Figure E8: Unpooled Treatment Effects on Information Outcomes

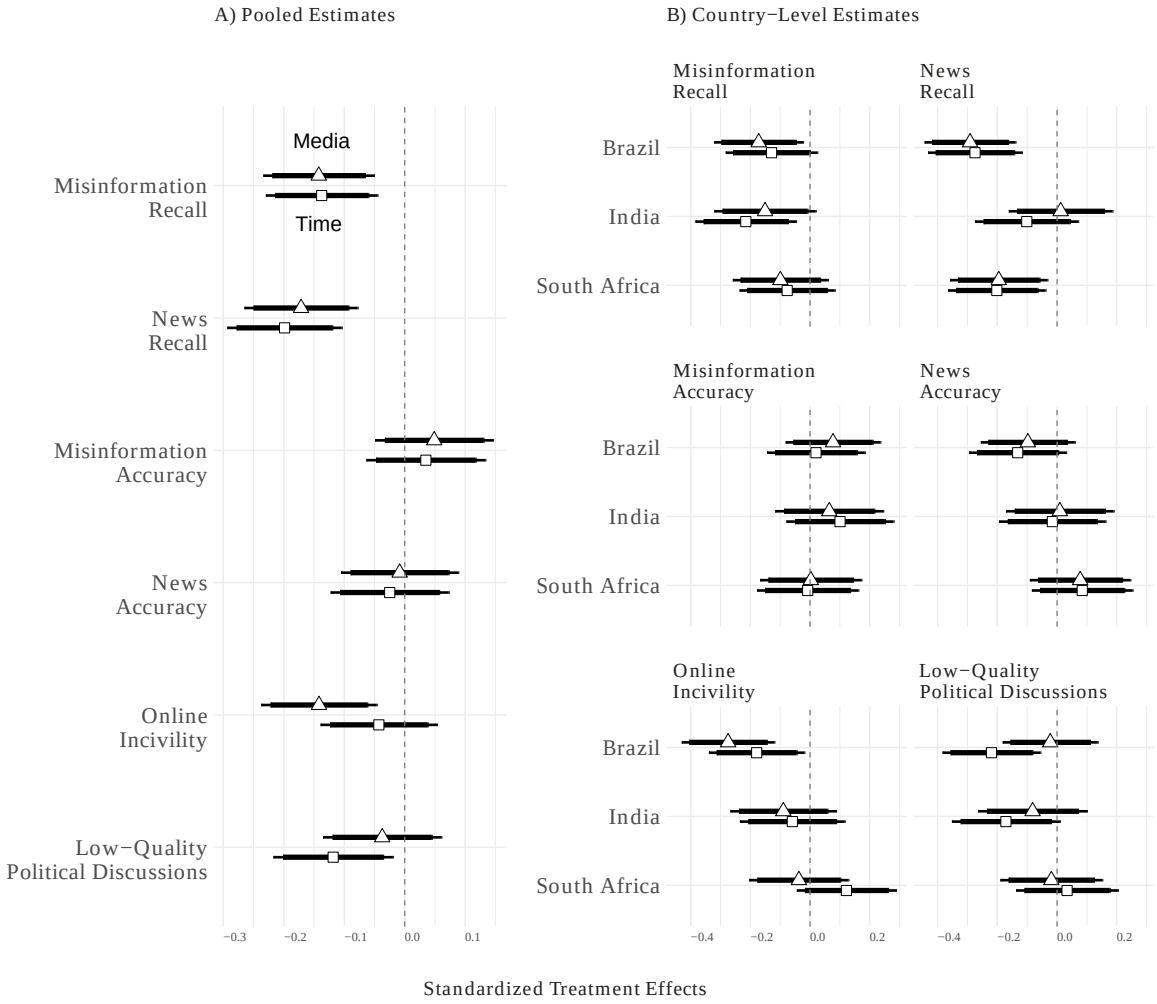


Figure E9: Unpooled Treatment Effects on Political Outcomes

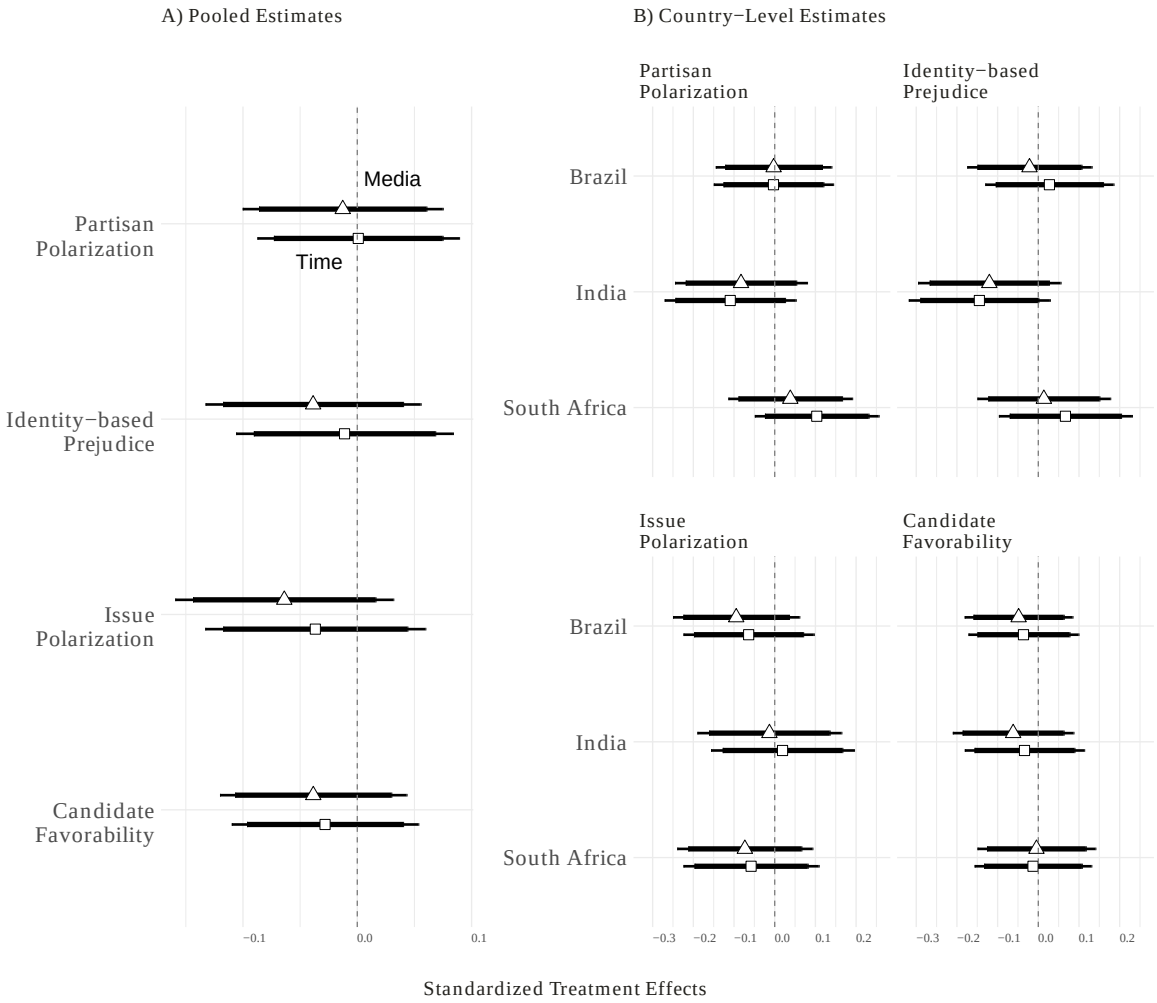
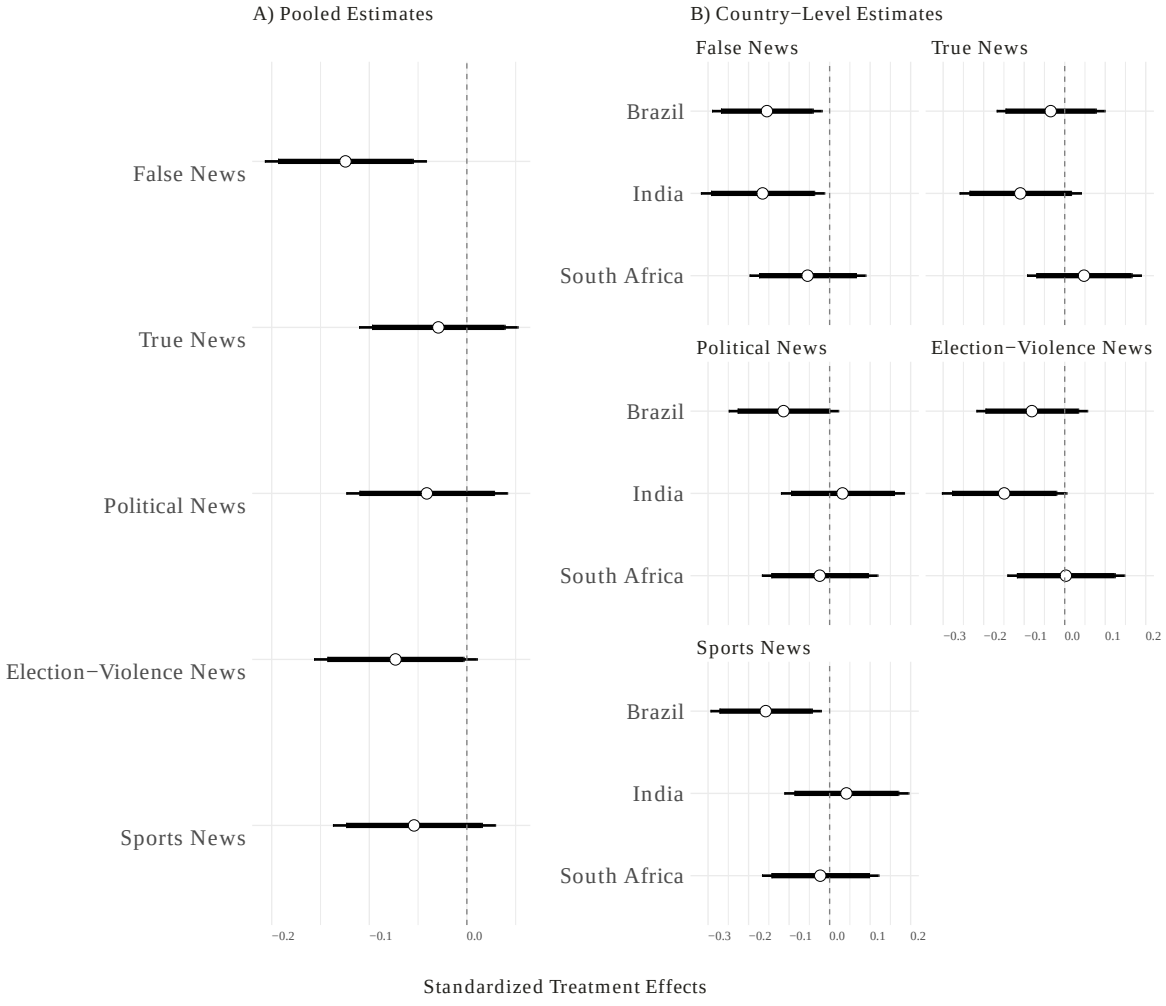


Figure E10: Unpooled Treatment Effects on Substitutes and Well-Being

E.2 Treatment Effects on Self-Reported Exposure to News

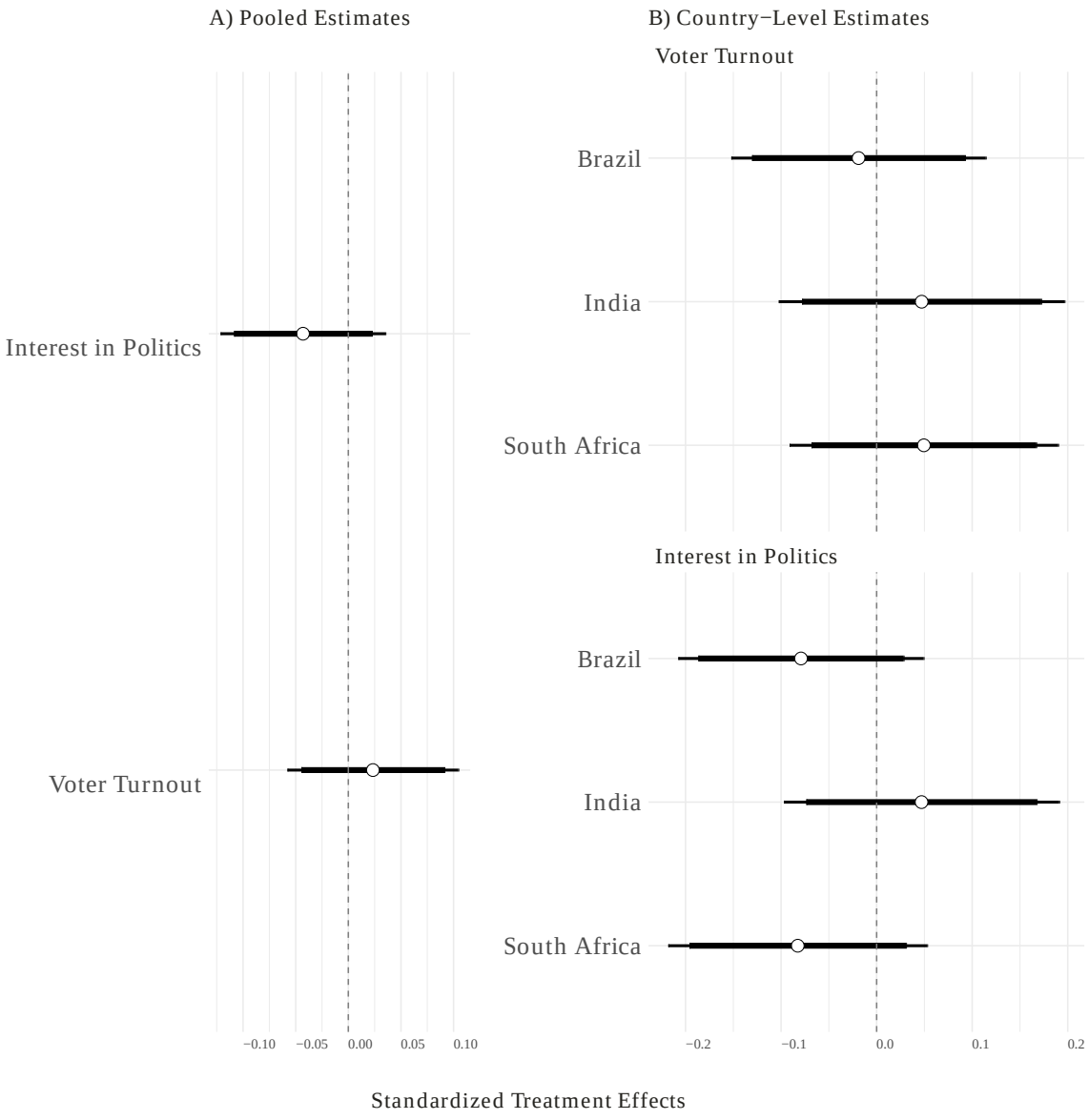
In addition to measuring changes in information consumption using recall of specific true and false rumors headlines, we also asked participants directly about the type of information they consumed during the weeks of the intervention. Specifically, we asked: "Thinking back over the past one month, how much did you see the following on social media?" for the items: (a) information you think is true; (b) information you think is false; (c) news about politics; (d) news about election related violence; and (e) news about sports. Responses were collected on a five-point scale ranging from "never" to "very frequently." These outcomes help us understand how participants perceive changes in their informational environment after reducing social media usage; they do not replace our pre-registered results using the headlines recall task. Figure E11 presents the results.

Figure E11: Treatment Effects on Self-Reported Exposure to News Content



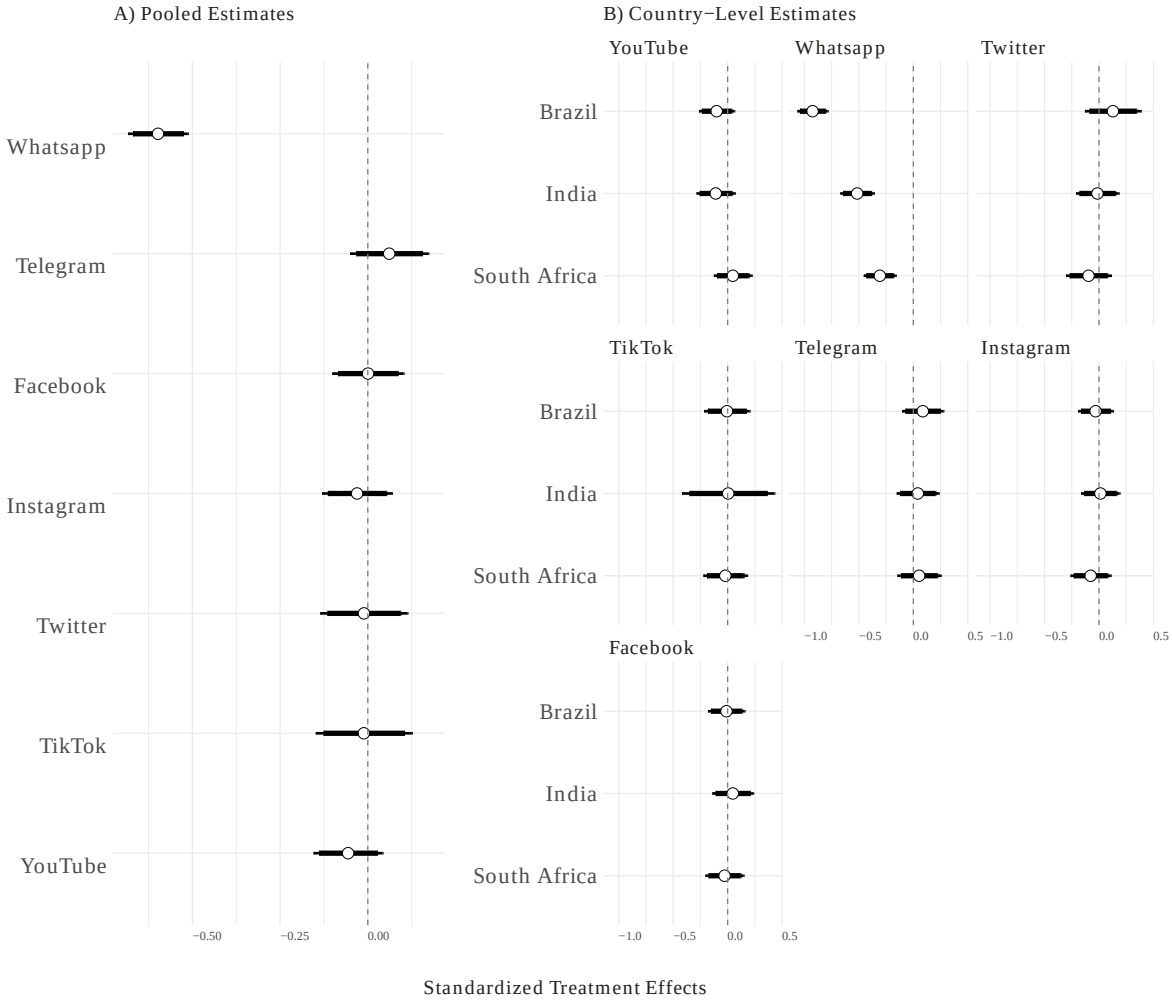
E.3 Political Interest and Voter Turnout

Figure E12: Treatment Effects on Political Interest and Voter Turnout



E.4 Substitution to Other Social Media Applications

Figure E13: Treatment Effects on Substitution to Distinct Social Media Platforms



E.5 Multiple Hypotheses Testing

In this section, we present the results reported in the paper for our pre-registered outcomes after correcting for multiple hypotheses testing. We use the Benjamini-Hochberg sharpened False Discovery Rate (FDR) adjustment for the potential for false discovery. We adjust the information outcomes by six hypotheses and the attitudinal outcomes by four hypotheses. Table E12 presents the results. Except for **H4**, all of our outcomes which have statistically significant effects at conven-

tional 95% confidence intervals remain so with adjusted p-values smaller than 0.05. With adjusted p-values, the treatment effects on exposure to low-quality political discussions is only significant at the 90% level.

Table E12: Unadjusted and FDR Adjusted P-Values Testing Each Hypothesis)

Hypotheses	Outcome	Unadjusted P-Value	FDR Adjusted P-Value
Information Outcomes			
H1a	Misinformation Recall	0.000	0.001
H1b	News Recall	0.000	0.000
H2a	Misinformation Accuracy	0.313	0.376
H2b	News Accuracy	0.648	0.648
H3	Online Incivility	0.022	0.044
H4	Low-Quality Political Discussions	0.049	0.074
Attitudinal Outcomes			
H5	Partisan Polarization	0.885	0.885
H6	Identity-based Prejudice	0.546	0.728
H7	Issue Polarization	0.195	0.635
H8	Candidate Favorability	0.317	0.635

Note: The unadjusted p-values are estimated using multilevel models for the pooled treatment effects with covariates selected via Lasso. For Information Outcomes, we adjust for 6 comparisons, while for the Attitudinal Outcomes, we adjust for 4 simultaneous comparisons. We use Benjamini-Hochberg sharpened False Discovery Rate (FDR) for adjustment.

E.6 Heterogenous Treatment Effects

In this section, we present pre-registered heterogeneous treatment effects for all the main outcomes discussed in the paper. Results are presented in the regression tables below as well as Figure E14.

Table E13: Regression Models: Heterogeneous eEffects Conditional on Digital Literacy

	Misinformation Exposure	News Exposure	Misinformation Beliefs	News Knowledge	Online Incivility	Low-Quality Political Discussions	Partisan Polarization	Identity-based Prejudice	Issue Polarization	Candidate Favorability	Subjective Well-Being
Treatment	-0.148+	-0.171*	0.127	0.005	-0.011	0.216	-0.179	-0.037	-0.224	-0.226+	0.485+
	(0.078)	(0.085)	(0.080)	(0.087)	(0.128)	(0.182)	(0.153)	(0.144)	(0.219)	(0.121)	(0.259)
Digital Literacy	-0.093***	-0.009	0.138***	0.043+	0.027	0.066	-0.041	-0.025	-0.040	-0.063+	0.051
	(0.022)	(0.024)	(0.022)	(0.025)	(0.036)	(0.051)	(0.044)	(0.040)	(0.062)	(0.034)	(0.072)
Treatment x Digital Literacy	-0.004	-0.028	-0.041	-0.013	-0.062	-0.186**	0.080	-0.017	0.032	0.076	0.006
	(0.030)	(0.033)	(0.031)	(0.034)	(0.050)	(0.071)	(0.060)	(0.056)	(0.085)	(0.047)	(0.101)
Num.Obs.	2222	2222	2222	2220	2222	2214	2220	2222	2222	2121	2137
R2 Cond.	0.142	0.205	0.066	0.066			0.279		0.081	0.348	0.019
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note:
Robust standard errors in Parentheses. All models use a multilevel estimation with random intercepts at the country level, and a list of covariates selected via Lasso for each outcome.

Table E14: Regression Models: Heterogeneous effects Conditional on Age

	Misinformation Exposure	News Exposure	Misinformation Beliefs	News Knowledge	Online Incivility	Low-Quality Political Discussions	Partisan Polarization	Identity-based Prejudice	Issue Polarization	Candidate Favorability	Subjective Well-Being
Treatment	-0.281*	-0.319**	0.005	-0.021	0.064	0.037	-0.260	-0.100	-0.284	-0.147	0.690+
	(0.110)	(0.120)	(0.114)	(0.123)	(0.180)	(0.257)	(0.216)	(0.204)	(0.310)	(0.171)	(0.368)
Age	0.002	0.049	0.025	-0.019	-0.249***	0.006	-0.050	-0.074	0.072	0.161***	0.217*
	(0.031)	(0.034)	(0.032)	(0.034)	(0.050)	(0.071)	(0.061)	(0.056)	(0.086)	(0.048)	(0.101)
Treatment x Age	0.050	0.037	0.018	0.000	-0.087	-0.092	0.107	0.011	0.051	0.034	-0.083
	(0.043)	(0.047)	(0.045)	(0.048)	(0.070)	(0.100)	(0.084)	(0.079)	(0.121)	(0.067)	(0.144)
Num.Obs.	2221	2221	2221	2219	2221	2213	2219	2221	2221	2120	2136
R2 Cond.	0.133	0.206	0.047	0.065			0.279		0.081	0.347	0.019
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note:
Robust standard errors in Parentheses. All models use a multilevel estimation with random intercepts at the country level, and a list of covariates selected via Lasso for each outcome.

Table E15: Regression Models: Heterogeneous effects Conditional on Overall WhatsApp Usage

	Misinformation Exposure	News Exposure	Misinformation Beliefs	News Knowledge	Online Incivility	Low-Quality Political Discussions	Partisan Polarization	Identity-based Prejudice	Issue Polarization	Candidate Favorability	Subjective Well-Being
Treatment	0.212+	-0.079	-0.166	-0.418**	0.179	0.334	-0.236	0.024	-0.113	-0.010	0.679
	(0.125)	(0.137)	(0.130)	(0.140)	(0.205)	(0.291)	(0.245)	(0.231)	(0.352)	(0.194)	(0.417)
WhatsApp Usage	0.048*	-0.014	-0.007	-0.105***	0.130***	0.221***	0.000	0.039	0.000	0.022	-0.018
	(0.024)	(0.026)	(0.025)	(0.026)	(0.038)	(0.055)	(0.047)	(0.044)	(0.066)	(0.037)	(0.078)
Treatment x WhatsApp Usage	-0.103**	-0.042	0.059+	0.110**	-0.089+	-0.142+	0.062	-0.027	-0.013	-0.016	-0.049
	(0.032)	(0.035)	(0.033)	(0.036)	(0.053)	(0.075)	(0.063)	(0.059)	(0.090)	(0.050)	(0.107)
Num.Obs.	2222	2222	2222	2220	2222	2214	2220	2222	2222	2121	2137
R2 Cond.	0.138	0.205	0.048	0.068			0.279		0.081	0.348	0.019
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note:
Robust standard errors in Parentheses. All models use a multilevel estimation with random intercepts at the country level, and a list of covariates selected via Lasso for each outcome.

Table E16: Regression Models: Heterogeneous effects Conditional on WhatsApp Usage for News

	Misinformation Exposure	News Exposure	Misinformation Beliefs	News Knowledge	Online Incivility	Low-Quality Political Discussions	Partisan Polarization	Identity-based Prejudice	Issue Polarization	Candidate Favorability	Subjective Well-Being
Treatment	0.112	0.087	-0.056	-0.252	0.705*	0.778+	-0.260	-0.279	-0.151	-0.322	0.314
	(0.197)	(0.215)	(0.204)	(0.221)	(0.322)	(0.457)	(0.388)	(0.364)	(0.554)	(0.310)	(0.660)
WhatsApp for News	0.084**	0.112***	-0.049+	-0.008	0.156***	0.216**	-0.022	-0.012	0.108	-0.065	0.128
	(0.029)	(0.031)	(0.029)	(0.032)	(0.047)	(0.066)	(0.057)	(0.053)	(0.080)	(0.045)	(0.093)
Treatment x WhatsApp for News	-0.054	-0.063	0.020	0.046	-0.167**	-0.189*	0.049	0.040	-0.002	0.050	0.036
	(0.038)	(0.041)	(0.039)	(0.042)	(0.062)	(0.088)	(0.075)	(0.070)	(0.106)	(0.059)	(0.127)
Num.Obs.	2222	2222	2222	2220	2222	2214	2220	2222	2222	2121	2137
R2 Cond.	0.137	0.210	0.048	0.065			0.279		0.082	0.348	0.022
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note:
Robust standard errors in Parentheses. All models use a multilevel estimation with random intercepts at the country level, and a list of covariates selected via Lasso for each outcome.

Table E17: Regression Models: Heterogeneous effects Conditional on How Often Receive Political Content on WhatsApp

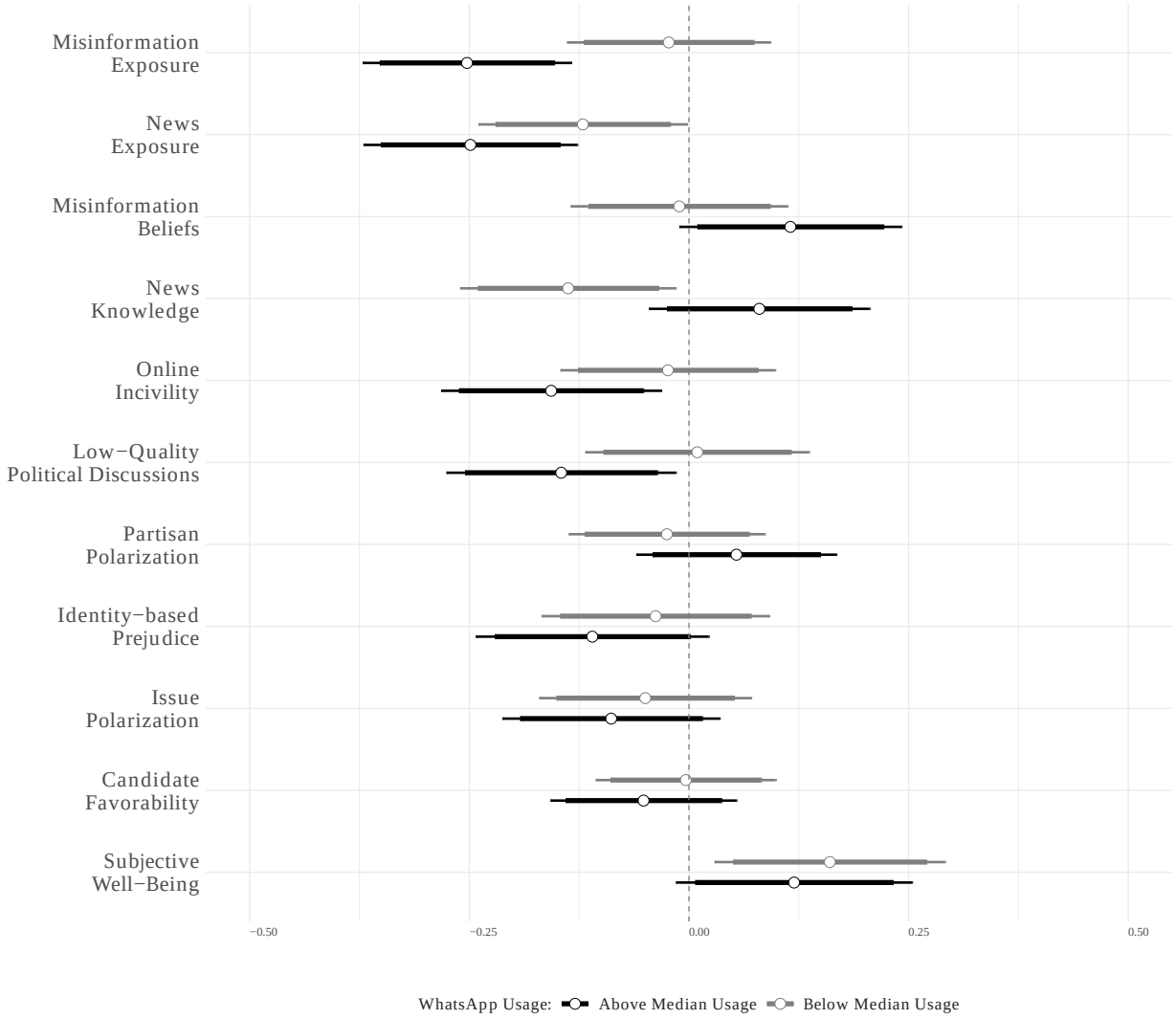
	Misinformation Exposure	News Exposure	Misinformation Beliefs	News Knowledge	Online Incivility	Low-Quality Political Discussions	Partisan Polarization	Identity-based Prejudice	Issue Polarization	Candidate Favorability	Subjective Well-Being
Treatment	0.016	-0.197	-0.122	-0.308+	0.482+	0.481	0.210	0.121	-0.072	0.050	0.442
	(0.167)	(0.182)	(0.172)	(0.186)	(0.271)	(0.386)	(0.326)	(0.308)	(0.468)	(0.263)	(0.556)
WhatsApp Political Content	0.065**	0.074**	-0.063*	0.012	0.185***	0.259***	0.047	0.042	0.128+	0.044	-0.024
	(0.025)	(0.027)	(0.025)	(0.027)	(0.039)	(0.056)	(0.049)	(0.045)	(0.068)	(0.038)	(0.079)
Treatment x WhatsApp Political Content	-0.038	-0.008	0.036	0.060	-0.133*	-0.141+	-0.046	-0.041	-0.020	-0.025	0.013
	(0.034)	(0.037)	(0.035)	(0.038)	(0.055)	(0.078)	(0.066)	(0.062)	(0.095)	(0.053)	(0.112)
Num.Obs.	2222	2222	2222	2220	2222	2214	2220	2222	2222	2121	2137
R2 Cond.	0.138	0.207	0.049	0.066			0.280		0.083	0.347	0.019
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note:

Note: Robust standard errors in Parentheses. All models use a multilevel estimation with random intercepts at the country level, and a list of covariates selected via Lasso for each outcome.

Figure E14: Treatment Effects Conditional on Low and High WhatsApp Usage



E.7 Cross-country Comparisons

Figures E15, E16, and E17 present average responses to key post-treatment and pre-treatment variables among untreated respondents in each country. In the main text, we consider how cross-country differences in baseline attitudes and exposure to content can help inform differences in treatment effects across countries. We also note that there are no discernible differences between our "time control group" and "media control group" with respect to these variables, assuaging any concerns one may have about pooling the control groups in our analyses.

Figure E15: Control group responses by country (headline recall and accuracy outcomes)

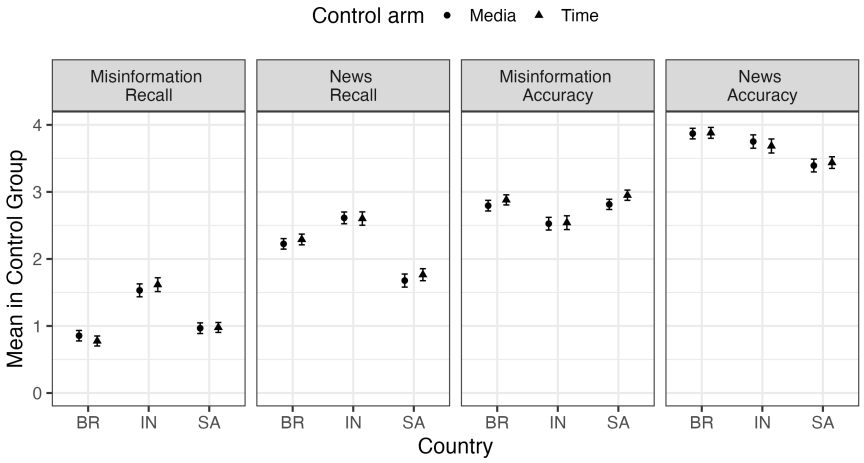


Figure E16: Control group responses by country (incivility and quality of discussions outcomes)

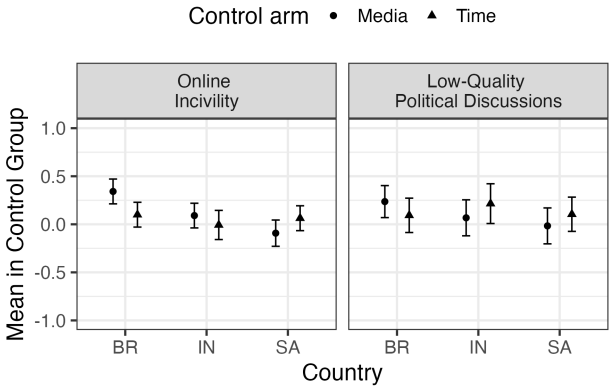
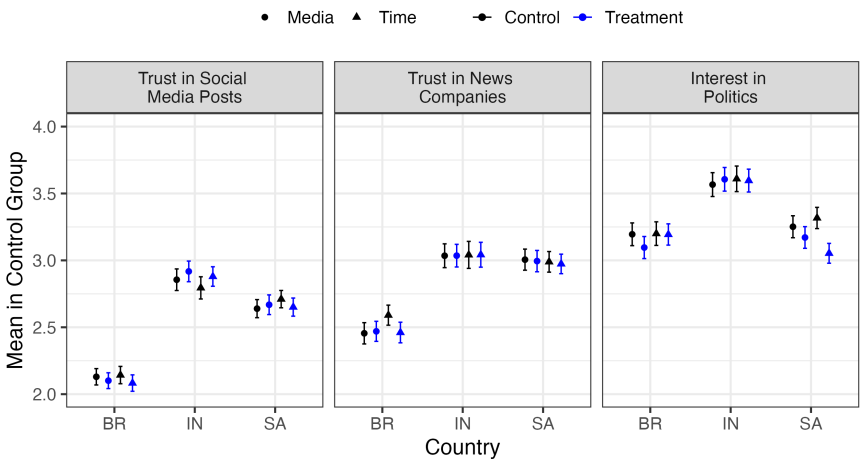


Figure E17: Control group responses by country (pre-treatment variables)

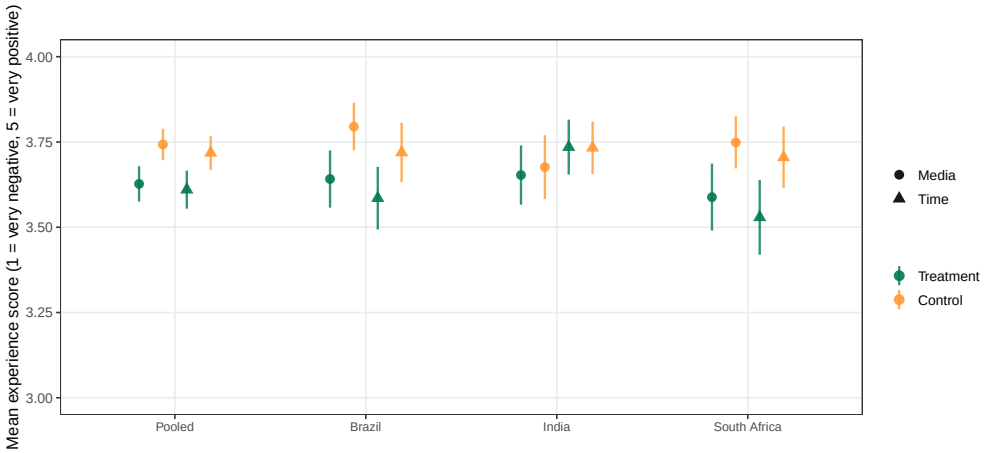


E.8 Participant Perspectives on Study Participation

In this section, we explore respondents' perspectives on participating in the study. We use data from two questions in the endline survey: a close-ended question on overall participation experience (sent to all respondents) and an open-ended question on the experience of limiting WhatsApp usage for four weeks (sent only to participants assigned to the treatment condition).

First, Figure E18 presents responses to the close-ended question, plotting the mean response by country and treatment condition. Across the different countries and conditions, the average response was above 3.5 (where 3 indicates neutral and 4 indicates positive), though treated individuals did report having a marginally less positive experience than those in the control – which is not surprising given the major lifestyle change the former undertook compared to the latter's experience of just completing a few short surveys. This provides a promising foundation when considering the scalability of such interventions. Though the intervention may have been disruptive, respondents generally did not find it to be unacceptable. As we discuss next, their open-ended responses corroborate this argument and reveal a rather nuanced picture, highlighting that the experience was rewarding and offered unexpected benefits for many participants even if it brought some difficulties with it.

Figure E18: Close-Ended Study Experience Responses, By Country and Condition



Each treated participant was further asked to narrate their experience deactivating from WhatsApp during the preceding four weeks.[17] Between 98% and 99.5% of respondents across the three coun-

[17]We did not ask control condition participants this question as we were specifically interested in what individuals

tries and two treatment types replied. In reading their responses, we observed the following themes come up repeatedly: the difficulty of refraining from using an app as deeply embedded in individuals' lives as WhatsApp is, the benefits of not seeing unnecessary or annoying content anymore as a function of having to prioritize what one views during their limited time on the app, and the advantage of suddenly having significantly more time to pursue other activities. For example, a South African participant wrote, "it was hard at first and felt like i was missing out on the rest of the day's interactions but i got used to it and started feeling like i had so much time to my hands and could do a lot more things." Similarly, a Brazilian user shared, "It was a big challenge, the first few days were the hardest, but I kept myself busy with other things, my boyfriend liked that I spoke to him more in person instead of on WhatsApp." In India, another respondent noted, "a lot of negative energy has lifted, i really felt this was a great move because whenever i used to consume whatsapp media most of it will be false news and negativity... since i haven't consumed it for a long time, i feel very positive i never came across anything offensive or false."

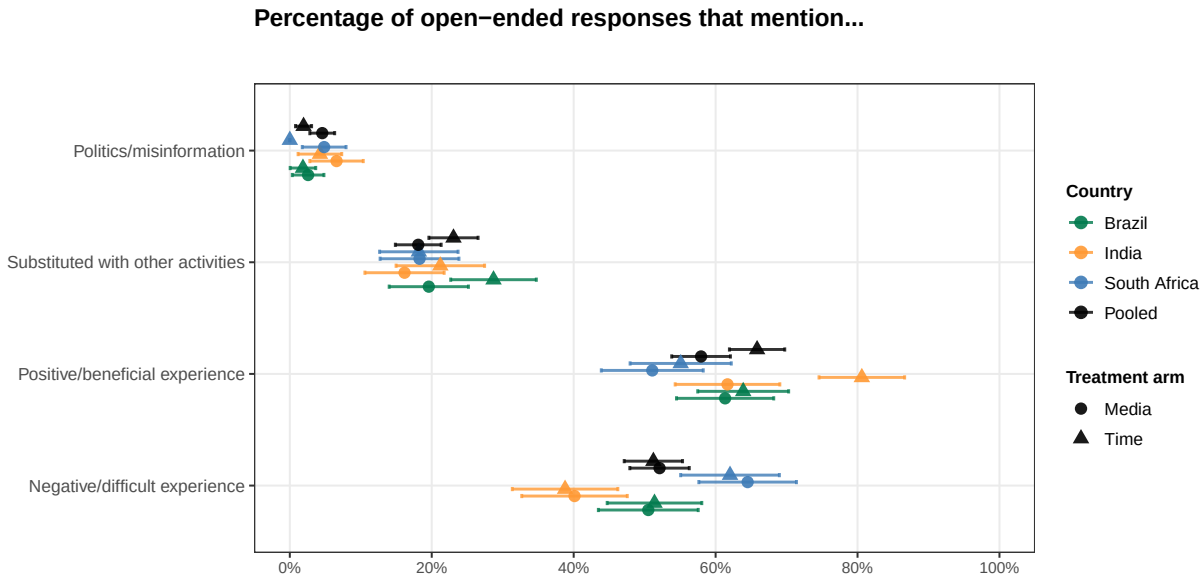
Accordingly, we used a large language model (specifically, Claude-3.5-Sonnet) to systematically classify the text responses across four dimensions. First, whether the respondent mentioned positive facets of the experience, such as enjoyment, ease of deactivation, reduced anxiety or stress as a result of being away from social media, or improved personal well-being. Second, whether the respondent reported increased anxiety, difficulty deactivating, missing out on important or interesting messages, or other negative feelings. Third, whether the respondent mentioned substituting WhatsApp with other activities like spending time with friends and family, watching television, pursuing hobbies, being more productive at work, or using other social media apps. Fourth, whether they explicitly mentioned politics, elections, or misinformation. Figure E19 shows the proportion of open-ended responses coded as "yes" in each category across the entire sample and by country, disaggregated by treatment type. We consider this analysis to be preliminary and exploratory, subject to additional tweaks to the algorithm as appropriate in future iterations of the paper.

Respondents often had both positive and negative takeaways. Overall, pooling across the three countries, we find that 64% of respondents reported positive experiences and 52% reported neg-

thought about the deactivation experience, rather than how the overall study experience varied by the type of experience.

ative experiences, suggesting that many participants had mixed feelings about the deactivation (and felt comfortable sharing these feelings candidly with the research team). Further, 21% mentioned substitution activities and 3% explicitly referenced politics or misinformation – without anybody being prompted to write about any of these themes.[18]

Figure E19: Classification of Open-Ended Study Experience Responses, By Country and Condition



[18]Disaggregating by country, we find that in Brazil, 63% of the respondents wrote about positive experiences while 51% mentioned negative experiences. Approximately 24% mentioned substituting WhatsApp with other activities, while only 2% explicitly referenced politics or misinformation. India showed the highest proportion of positive experiences at 71%, with 39% reporting negative experiences and 19% mentioning substitution activities. Political content was mentioned by 5% of Indian respondents, the highest rate among the three countries. We uncovered a somewhat different pattern in South Africa, where 53% mentioned positive experiences but 63% brought up negative experiences.