Misinformation Exposure Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil

Tiago Ventura*[1, 2], Rajeshwari Majumdar[2, 3], Jonathan Nagler[2, 3], and Joshua A. Tucker[2, 3]

[1]Georgetown University
[2]Center for Social Media and Politics (CSMaP), New York University
[3]Wilf Family Department of Politics, New York University

August 16, 2024

Abstract

In most advanced democracies, concerns about the spread of misinformation are typically associated with feed-based social media platforms like Twitter and Facebook. These platforms also account for the vast majority of research on the topic. However, in most of the world, particularly in Global South countries, misinformation often reaches citizens through social media messaging apps, particularly WhatsApp. To fill the resulting gap in the literature, we conducted a *multimedia deactivation* experiment to test the impact of reducing exposure to potential sources of misinformation on WhatsApp during the weeks leading up to the 2022 Presidential election in Brazil. We find that this intervention significantly reduced participants' exposure to false rumors circulating widely during the election. However, consistent with theories of mass media minimal effects, a short-term reduction in exposure to misinformation ahead of the election did not lead to significant changes in belief accuracy, political polarization, or well-being.

*Corresponding Author. E-mail: tv186@georgetown.edu

Introduction

The spread of misinformation on social media and its potential to shape citizens' attitudes and behavior has long been a source of concern across the world. Yet the vast majority of the academic scholarship about the prevalence of misinformation and the role of social media in its spread has been conducted in the context of advanced democracies (Vosoughi, Roy and Aral, 2018; Grinberg et al., 2019; Nyhan et al., 2023). While concerns about misinformation in Global South democracies have received considerable attention in the media and from local policymakers, there remains a scarcity of scholarly knowledge regarding the dynamics of misinformation spread and its causal effects on attitudes and behavior outside of the United States and Europe. The lack of research on settings outside of Western democracies severely affects the generalizability of scholarly evidence and the validity of policy recommendations, especially in contexts where the deleterious offline consequences of misinformation exposure may be more pronounced (Badrinathan and Chauchard, 2023*a*; Tucker et al., 2018; Mitchelstein and Boczkowski, 2021).

This gap in the literature is made more salient by the fact that the types of social media platforms that are predominantly used in the Global South are fundamentally different in nature to those used in the Global North. In the latter, misinformation spread and exposure are commonly linked to conventional feed-based platforms like Twitter and Facebook. However, in other regions of the world, these concerns have been more strongly associated with messaging apps, such as WeChat, Telegram, and, most prominently, WhatsApp (Newman et al., 2021; Resende, Melo, Reis, Vasconcelos, Almeida and Benevenuto, 2019; Machado et al., 2019; Batista Pereira et al., 2023; Valenzuela, Bachmann and Bargsted, 2021).

With close to 2 billion active users worldwide, WhatsApp, an encrypted messaging app that allows both one-to-one and group communications, is the leading social media platform in most Global South countries, including India, Brazil, Nigeria, and South Africa. Users rely on WhatsApp to communicate with friends, conduct business, and consume news, including content about politics and elections (Rossini, Stromer-Galley, Baptista and Veiga de Oliveira, 2021). WhatsApp groups in particular have become a powerful propaganda and organizational tool and are heavily used to mobilize citizens around social, political, and identity issues (Chauchard

and Garimella, 2022; Gil de Zúñiga, Ardèvol-Abreu and Casero-Ripollés, 2021). Some have even argued that WhatsApp has had serious downstream effects on offline violence against minority groups (Saha et al., 2021; Banaji et al., 2019) and on voting behavior (Tardáguila, Benevenuto and Ortellado, 2018; Mello, 2020). Yet, no studies have measured the causal effects of WhatsApp usage on exposure to misinformation and downstream effects on users' attitudes and behaviors. As a result, there is a large gap between popular accounts of the role of social media messaging apps in fostering beliefs in misinformation and academic evidence related to these issues.

To address these gaps in the literature, our study focuses on identifying the causal effects of the most heavily used social media messaging app in the world (WhatsApp) on exposure to online rumors and its downstream effects on attitudes. To do so, we deploy a field experiment with WhatsApp users in the context of the 2022 general election in Brazil. Modeled after previous studies with Facebook users (Asimovic et al., 2021; Allcott et al., 2020), we randomly assign users to a *Multimedia WhatsApp Deactivation*, in which we ask treated participants to turn off their automatic download of multimedia – that is, images, videos, documents, and audio – received on WhatsApp and further incentivize them to not access any media during the three weeks leading up to the general election on October 2, 2022. Turning off automatic downloads on WhatsApp introduces substantial friction to consuming content on the platform without drastically changing users' daily habits. By blurring all multimedia received in personal conversations and group-based chats, this setting requires users to undertake an extra tap to view the content; users in our treatment condition are asked to refrain from doing that. Of course, evaluating compliance with such tasks in a completely online setting poses an interesting design challenge, which we address by asking participants to send us weekly screenshots of their WhatsApp storage statistics. In this way, we are able to precisely measure compliance with the multimedia deactivation by examining the amount of data consumed by each respondent during the experimental period.

We focus on *multimedia deactivation* for two primary reasons. First, the dynamics of information propagation on WhatsApp significantly differ from feed-based platforms. While the latter relies on algorithmic news feeds and user social graphs for information sharing, content

propagation on WhatsApp depends more heavily on users' decisions to forward content, both in group settings and in one-to-one chats. Therefore, without creators producing content for their followers, the most viral information on WhatsApp travels across chats in a quasi-anonymous format, lacking any metadata, and is often crafted for easy distribution across different groups and chat conversations. As a consequence, instead of text-based news articles and posts, it is easy-to-share multimedia content (such as videos, images, audio, and GIFs) that dominates WhatsApp's informational environment, and to a large degree, represents most of the misinformation circulating in the platform (Burgos, 2019; Avelar, 2019; Resende, Melo, Reis, Vasconcelos, Almeida and Benevenuto, 2019; Machado et al., 2019; Garimella and Eckles, 2020). Second, considering how embedded WhatsApp is in users' personal and professional lives in a way 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), a study that assigns subjects to a complete deactivation would be problematic both theoretically and empirically, as the subset of people willing to deactivate WhatsApp entirely would be a very distinct population from our population of interest (that is, regular users of WhatsApp).

Three days after election day, that is, after three weeks of treatment, we fielded a post-treatment survey to test four pre-registered hypotheses of interest. First, we test our hypothesis about the effects of the multimedia deactivation on exposure to true and false information that circulated widely on social media during the weeks of the presidential race. This allows us to causally assess the extant descriptive evidence pointing towards the prevalence of misinformation on WhatsApp (Resende, Melo, Reis, Vasconcelos, Almeida and Benevenuto, 2019; Freitas Melo et al., 2019; Machado et al., 2019; Garimella and Tyson, 2018; Garimella and Eckles, 2020). Second, building upon the well-established idea of the "illusory truth effect" hypothesis, which predicts that repeated exposure to a false statement increases its perceived accuracy (Fazio et al., 2015; Dechêne et al., 2010; Pennycook, Cannon and Rand, 2018), we test the effects of the deactivation and the expected reduction in exposure to misinformation on participants' capacity to correctly identify false rumors and true news stories. Lastly, we test for downstream effects on political and affective polarization and subjective well-being, outcomes that have been

examined extensively in the social media literature more broadly (Bail et al., 2018; Settle, 2018; Asimovic et al., 2021; Allcott et al., 2020).

Our results provide causal evidence of a substantive reduction in exposure to misinformation. After the deactivation, users report a statistically significant reduction in exposure to false rumors that circulated widely in the pre-election weeks. This reduction occurs at a higher rate when compared to changes in exposure to true news stories. However, despite these substantive changes in misinformation exposure, we fail to capture attitudinal changes in a) accuracy judgments for the same set of false and true stories, b) political and affective polarization, or c) self-reported subjective well-being. These findings are consistent with the media minimal-effects tradition on political attitudes (Lazarsfeld, Berelson and Gaudet, 1968; Bennett and Iyengar, 2008). Our findings also challenge arguments connecting exposure to news and misinformation rumors with belief formation in a purely mechanical manner (Pennycook, Cannon and Rand, 2018). The null findings in regard to three of our four primary hypotheses dovetail with recent field experiments conducted on Facebook that measure the effects of news exposure on a large set of attitudinal outcomes (Guess et al., 2023; Guess, Stroud and Tucker, 2023; Nyhan et al., 2023).

Additionally, we also test three pre-registered heterogeneous treatment effects for the covariates age, digital literacy, and use of WhatsApp for consuming political news. Our results suggest an interesting heterogeneity among participants who report frequently using WhatsApp to receive and share political news. Consistent with our pre-registered expectations, we find that reducing exposure to misinformation among users who report receiving and sharing political news frequently on WhatsApp improves their capacity to identify false rumors. At the same time, subjects who do not use WhatsApp to share and receive news about politics frequently become worse at identifying true and false rumors after the deactivation. We do not observe heterogeneous effects by age or digital literacy.

Our field experiment contributes to the thorny issue of identifying the causal effects of social media usage on information consumption and political attitudes. To date, we are only aware of prior studies identifying the causal effects of randomized deactivations of Facebook users (Han-

ley, Watt and Coventry, 2019; Vanman, Baker and Tobin, 2018; Allcott et al., 2020; Asimovic et al., 2021) and with a strong focus on the US case. In addition, our study also pushes the frontier of research on the use of encrypted messaging applications for political communication. Part of the challenge in studying WhatsApp and other social media messaging apps stems from data limitations that are inherent to an encrypted application. As a consequence, the vast majority of the emerging academic literature on WhatsApp has primarily focused on proposing strategies to collect data and, by doing so, providing a valuable description of the content that circulates through the platform (Avelar, 2019; Burgos, 2019; Garimella and Eckles, 2020; Chauchard and Garimella, 2022; Machado et al., 2019; Resende, Melo, Reis, Vasconcelos, Almeida and Benevenuto, 2019). While these efforts are valuable and inform our research, descriptive findings do not allow us to make causal claims about the effects of social media platforms. In addition, some recent work has focused on deploying interventions to counter beliefs on misinformation through WhatsApp, particularly exposing WhatsApp users to fact-checking content (Badrinathan and Chauchard, 2023*b*; Bowles, Larreguy and Liu, 2020). Our study differs from those by focusing on the direct causal effects of WhatsApp usage on misinformation-related outcomes and political attitudes, rather than using WhatsApp as a tool for misinformation interventions. We are the first to do so, and our findings present a rather complex picture of the role of WhatsApp usage in shaping accuracy perceptions and political attitudes among voters.

Misinformation, Politics, and the Brazilian Elections

Our study was conducted in Brazil prior to the general elections held on 2 October 2022 to elect the President, the National Congress, governors, and legislative assemblies at the state level. Due to institutional and historical reasons, Presidential races, to a large degree, dominate voters' attention and organize most of the political cleavages. This top-down dynamic was particularly true in the 2022 election, in which the incumbent President, Jair Bolsonaro, ran against the former President, Luiz Inacio Lula da Silva, in what has been considered the most polarized election in Brazilian democratic history. In a slim runoff election, Lula received 50.90% of the votes, becoming the first politician to defeat an incumbent President running for reelection. As expected, given his previous claims of electoral fraud, Bolsonaro and his Party

challenged the results, while his supporters have taken the streets claiming, among other things, a political intervention from the Armed forces. The Superior Electoral Court (TSE) ratified Lula's victory and is acknowledged by Brazil's leading politicians and international allies.

While Bolsonaro remained silent without recognizing his defeat, the Brazilian Superior Electoral Court (TSE) ratified Lula's victory. On January 8th, thousands of his supporters stormed Brazil's Congress, the Supreme Court, and presidential offices in anti-democratic scenes resembling the January 6th attacks in the United States closely. As usual, the role of misinformation received through social media in these protests and in fostering beliefs for the false claims sponsored by the rioters have occupied a prominent space in the media and have many times been argued as the cause for these anti-democratic acts.[1]

Online misinformation and its effects on politics have been a central theme in Brazilian politics for the last several years, especially since the 2018 elections and during the COVID-19 pandemic. According to a major worldwide survey of internet users published by the Reuters Institute (Newman et al., 2021), 83% of the Brazilian population reports receiving their news online, with 63% getting their news primarily through social media. Brazilians are among the most active users of platforms like Facebook, WhatsApp, and Twitter globally. WhatsApp is the most widely used social media application for all purposes in Brazil and is used by 43% of the general population for news consumption, only slightly behind Facebook (47%). In addition, Newman et al. (2021) report that in 20221 84% of the population in Brazil were concerned about *what is real and what is fake on the internet* – the highest among the 40 countries included in the Reuters Report – and that WhatsApp is the social media application where users report seeing most misinformation. Our focus on exposure to misinformation on WhatsApp is driven not only by the general lack of research on the causal impacts of WhatsApp usage but primarily by the salience of this issue in Brazil and most other Global South democracies.

[1]For examples in the international media, see the following links: https://nypost.com/2023/01/12/brazil-rioters-plotted-riot-on-facebook-youtube-and-twitter/, https://techpolicy.press/election-disinformation-and-the-violence-in-brazil/, and https://www.nytimes.com/2023/01/09/technology/brazil-riots-jan-6-misinformation-social-media.html

Online Misinformation, Illusory Truth Effects, Beliefs and Polarization

The complex role of social media and the internet in facilitating the spread of misinformation and magnifying its global consequences in a rapidly evolving digital information environment has become a prominent concern among policymakers, academic experts, and media (Lorenz-Spreen et al., 2022; Tucker et al., 2018). The arguments connecting social media usage to increased exposure to online misinformation take many forms. A more supply-side-based argument states that social media dramatically reduced the gatekeeping capacity of traditional reputable outlets to filter the circulation of news content, leading to a fragmentation of the media market and a reduction in the quality of online news sources (Aruguete, Calvo and Ventura, 2021). In this context of higher fragmentation and faster propagation, scholars argue that social media facilitates users' exposure to online rumors which are more likely to be shared by communities of like-minded partisans (Del Vicario et al., 2016; Nyhan et al., 2023), generating downstream effects on belief in misinformation rumors, levels of polarization, and outgroup animosity (Rathje, Van Bavel and Van Der Linden, 2021). These effects are magnified by network dynamics of social media activation, with previous research showing evidence that false information has a faster diffusion cascade when compared to factual statements on social media (Del Vicario et al., 2016; Vosoughi, Roy and Aral, 2018) and that more ideologically extreme users share highly partisan content faster on social media (Arugute, Calvo and Ventura, 2022).

The vast majority of this research has focused on feed-based platforms. When looking at social media messaging apps, such as WhatsApp, specific affordances of these platforms introduce novel challenges to understanding the spread of online misinformation. In the absence of news feeds and news accounts from major communication networks on the platforms, the circulation of information — true and false — on WhatsApp depends heavily on users' online and offline networks, and on their choices to share content, particularly on WhatsApp groups. As a consequence, sharing mechanisms and cascading dynamics on WhatsApp differ greatly from feed-based platforms. In the latter, information is often associated with content producers, such as journalists, news organizations, politicians, and influencers, and is often delivered through text-based content, such as news articles, posts, and comments. On WhatsApp, content shared

is not easily traceable to a source, as forwarded information does not come with any metadata, and is often produced in multimedia formats that can be easily shared across distinct groups and chat conversations (Rossini, Stromer-Galley, Baptista and Veiga de Oliveira, 2021).

Given these platform affordances, multimedia content that is easy-to-share and quasi-anonymous has become a crucial part of WhatsApp's informational environment. More importantly, due to the challenges of tracing the source of the information, scholarly research has shown that such multimedia content has become the primary format of false and misleading content online (Resende, Melo, Sousa, Messias, Vasconcelos, Almeida and Benevenuto, 2019; Garimella and Eckles, 2020; Machado et al., 2019). For example, in India, Garimella and Eckles (2020) show that 13% of all images shared on a set of political groups contained misinformation, and Tardáguila, Benevenuto and Ortellado (2018) estimated that roughly half of all images circulating in political groups in the weeks before the 2018 Presidential election in Brazil were likely altered or distorted to convey false information. Research using survey data to understand user habits and information consumption on WhatsApp has reached similar conclusions (Rossini, Stromer-Galley, Baptista and Veiga de Oliveira, 2021; Rossini, Baptista, de Oliveira and Stromer-Galley, 2021). This heavy reliance on multimedia content is also aggravated by WhatsApp users' demographic composition; large swatches of its user base in developing countries exhibit low levels of education and digital literacy (Badrinathan and Chauchard, 2023*a*).

Building on these unique features of social media messaging apps, our deactivation design focuses on restricting channels that allegedly represent the main modes through which misinformation travels on WhatsApp: multimedia content. In this vein, our first hypothesis[2] assesses experimentally how deactivating users' consumption of multimedia on WhatsApp shapes exposure to false rumors that circulated in the weeks prior to the election:

> **Hypothesis 1:** Deactivated users will report lower levels of exposure to misinformation compared to those who use WhatsApp as usual.

Our second hypothesis considers the effects of exposure to misinformation on beliefs. A

[2]To provide full context and transparency, it should be noted that we switched the ordering of Hypotheses 1 and 2 between our pre-analysis plan (PAP) and the present manuscript. The content of the hypotheses remain unchanged. See section 14 in the SIF appendix for information on deviations from the PAP.

well-established literature in cognitive psychology uses the concept of "illusory truth effects" to demonstrate that prior exposure to a blatantly false statement increases perceptions of accuracy about political and non-political issues (see Dechêne et al. 2010 for a meta-analysis of the concept). In this argument, repeated exposure to falsehoods affects the ease with which false statements are processed and, therefore, increases their perceived accuracy. Recent work by Pennycook, Cannon and Rand (2018) expands the idea of the "illusory truth effect" to the case of online political information, including misinformation. According to their argument, online environments facilitate the spread of misinformation and help incubate among the population beliefs in false news stories as a consequence of repeated online exposure.

Our study aims to assess the "illusory truth effect" as a psychological mechanism. While evidence for Pennycook, Cannon and Rand (2018)'s argument has been found using internally valid survey experiments, evidence from more naturalistic interventions with high ecological validity is scarce.[3] We are among the first to intervene in a naturalistic setting to assess the downstream effects of reducing exposure to false rumors on belief in misinformation. In addition, our design imposes a hard test of the role of directional motivated reasoning as a scope condition for the role of "illusory truth effects" by intervening in a highly polarized political environment. Following this argument, we expect that:

> **Hypothesis 2a:** Deactivated users will be more likely to identify false rumors as false compared to those who use WhatsApp as usual.

However, outside of the issue of online misinformation, social media has become a central channel for citizens to encounter and engage with news and political information around the world in recent years (Newman et al., 2021) . More importantly, evidence indicates that social media usage exerts a substantial influence on news knowledge in today's media landscape, for example, fostering information consumption through trusted social connections (Bode, 2016), facilitating learning, access, and engagement with political information (Park and Gil de Zúñiga, 2020), or simply through incidental exposure to news and politics on social media sites (Nanz and Matthes, 2022). In the Global South, social media messaging apps have achieved such

[3]See Guess et al. (2021) for a similar design with a focus on web-browsing consumption.

prominence as an informational tool, and previous research shows that WhatsApp has become a primary channel for citizens to learn about politics and breaking news (Valenzuela, Bachmann and Bargsted, 2021; Rossini, Baptista, de Oliveira and Stromer-Galley, 2021). Considering that our intervention generates a short-term reduction in overall information exposure, including true content, we expect:

> **Hypothesis 2b:** Deactivated users will be less likely to identify true news stories as true compared to those who use WhatsApp as usual.

Our third hypothesis engages with the open debate about the effects of digital media, particularly exposure to online misinformation, on heightened levels of political and affective polarization around the world. Heightened levels of polarization have become a striking feature of contemporary politics, and the consolidation of social media in today's informational environment has been widely cited as a driving force behind this trend (Flaxman, Goel and Rao, 2016; Bail et al., 2018; Settle, 2018). Digital media's role in amplifying the dissemination of misinformation is widely discussed as a pivotal factor in exacerbating this trend (Flaxman, Goel and Rao, 2016; Del Vicario et al., 2016). Online misinformation tends to be produced, consumed, and shared along partisan lines, with political elites producing online misinformation and highly motivated partisan users acting as super-spreaders of these false narratives online. In this partisan-motivated story of misinformation consumption and production, fake news represents the extreme ends of a partisan news continuum that can distort ingroup consumers' perceptions of reality, increase their misperceptions about others who hold different political preferences, and deepen ideological divides. This theory is largely consistent with the empirical evidence from recent scholarly work (Rathje, Van Bavel and Van Der Linden, 2021; Suhay, Bello-Pardo and Maurer, 2018; Settle, 2018)

While previous scholarly work has shown that correcting misinformation beliefs and misperceptions about outgroup voters could reduce levels of polarization (Druckman et al., 2023), in line with previous deactivation studies (Asimovic et al., 2021; Allcott et al., 2020), our intervention assesses the causal effect a short-term reduction on exposure to online misinformation, and more broadly WhatsApp usage, on polarization. We focus on four common dimensions

in the polarization literature: misperceptions of ideological polarization, affective polarization, social polarization, and issue polarization (see SI Appendix Table 2 for the operationalization of our four measures). We expect:

> **Hypothesis 3:** Deactivated users will display lower levels of polarization compared to those who use WhatsApp as usual.

Lastly, social media adoption is often theorized to affect well-being. The existing literature argues that social media usage can lead to harmful social comparisons, replace well-being-boosting activities (such as in-person socializing), and exacerbate feelings of loneliness (Kross et al., 2021; Twenge and Campbell, 2018). While these arguments are supported with correlational evidence, recent randomized controlled trials, including social media deactivation studies and meta-analyses, have found mixed results for how reducing social media usage affects self-reported well-being measures (Allcott et al., 2020; Asimovic et al., 2021; Vanman, Baker and Tobin, 2018). Since our intervention might also lead to an overall reduction in WhatsApp usage among the treated participants, we pre-registered a secondary hypothesis about the effects of social media usage on individuals' subjective well-being:

> **Hypothesis 4:** Deactivated users will display higher levels of subjective well-being compared to those who use WhatsApp as usual.

Research Design

We conducted our experiment during the 2022 Brazilian Presidential election, starting on September 15 and ending on October 5, three days after the first round of the election. Figure 1 presents a stylized description of the experimental design.[4] We discuss the different stages of the design in further detail below.

Survey Recruitment: We used Facebook Advertisements to recruit participants for the experiment. The ads ran from September 8 to September 15 (2022). SI Appendix Section 1 presents a preview of the ads, which were displayed to adults currently accessing Facebook from Brazil.

[4]This research was approved by the [REDACTED]'s IRB (protocol IRB-FY2022-6727).

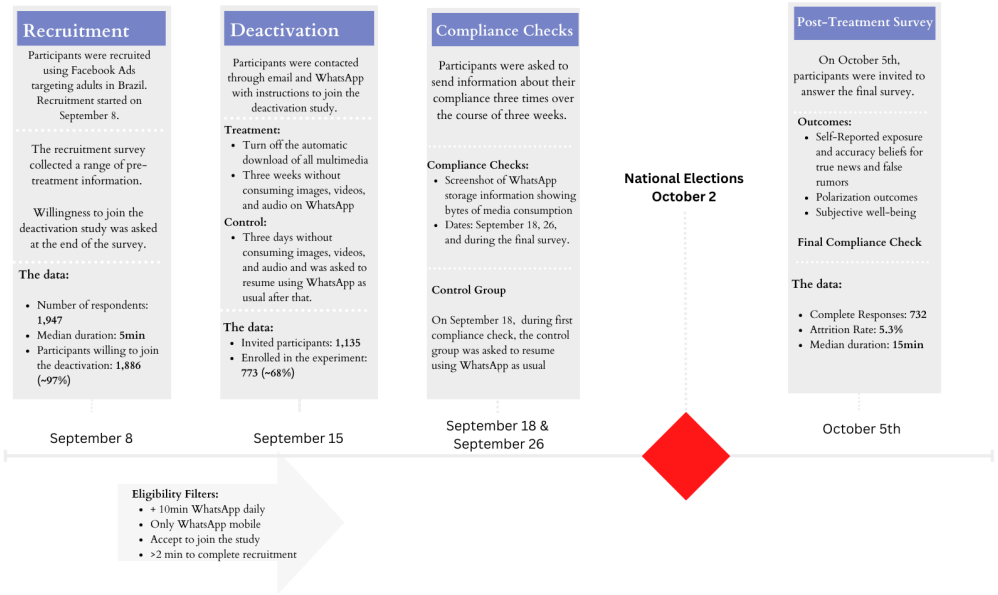


Figure 1 Overview of the WhatsApp Multimedia Deactivation Experiment

The ads directed participants from Facebook to a Qualtrics survey, in which participants answered a five-minute baseline survey with a battery of questions about their demographics and social media habits, including WhatsApp usage. In addition, at the end of the same questionnaire, we provided a detailed explanation of the experiment and asked participants about their willingness to take part in the study. The English translation for the invitation to join is provided in SI Appendix Section 1. Out of the respondents who completed the pre-treatment survey (N=1947), we selected participants based on three criteria: i) agreement to participate in the deactivation study; ii) self-reported WhatsApp usage of more than 10 minutes per day; iii) self-reported WhatsApp usage on mobile only (thus excluding WhatsApp desktop users).[5] The decision to select only participants who use WhatsApp mobile is a constraint stemming from our strategy to monitor compliance with the experiment, which relies on screenshots from participants' mobile storage settings.[6] Out of the entire pre-treatment sample, 95% of the re-

[5] As pre-registered, we also screened out participants who completed the entire baseline survey in less than 2 minutes, with the purpose of removing individuals racing through the questionnaire.

[6] Different from feed-based platforms, WhatsApp requires users to have a mobile account to access the app. Therefore, our decision to restrict the sample to mobile users hardly affects the composition of the sample

spondents reported using WhatsApp for more than 10 minutes per day, 75% reported using only WhatsApp mobile, and 97% of those of the participants who completed the pre-treatment survey agreed to join our study.

Treatment Assignment: On September 15, we invited 1,135 participants who passed the filters described above to take part in the deactivation experiment. We sent the invitations to their personal email addresses and WhatsApp numbers. Out of the individuals invited, 773 (68%) chose to enroll in the experiment. We block randomized treatment assignments across age, education, and gender into two distinct groups: control and treatment. The treatment group was asked to turn off their automatic downloads of multimedia on WhatsApp AND to not consume any media on WhatsApp for three weeks. The control group was asked not to consume any media on WhatsApp for three *days* (as opposed to three *weeks* for the treatment group). Specifically, the treatment group was instructed to: i) turn off the automatic download of media on WhatsApp; ii) upload a screenshot of their WhatsApp settings with automatic downloads disabled[7]; and iii) not view any multimedia content for three weeks.[8] The control group did not receive (i) or (ii), and their instructions for (iii) only asked them to not consume any media for three days. All users were offered the same financial incentive to participate[9], regardless of their treatment assignment.

A primary challenge in online, multi-wave experiments is attrition, particularly when that attrition may occur differentially across the treatment and control conditions. In our case, there are two potential sources of missingness: attrition among the participants we initially invited to the experiment and attrition among the participants who started the experiment but did not complete the post-treatment survey. Because of the strong dosage of the treatment intervention (a three-week deactivation), we were particularly worried about the former. To mitigate these

[7]On WhatsApp, users have the option to enable or disable the automatic download of media to their phones. When this feature is enabled, media received on WhatsApp opens directly in chat windows. When this feature is disabled, participants see blurred thumbnails, which they must click on to download and thereby view the content.

[8]These instructions came with detailed directions on how to navigate through these various directions. The full wording for these instructions is provided in the SI Appendix, Section 1.3

[9]We offered participants a total compensation of 150 Brazilian Reais (approximately 30 USD) for their participation in the study.

concerns, we consciously adjusted our design to reduce the experiment's costs for participants and to screen out those who would be unlikely to take the treatment regardless of whether it is offered (*never-takers*) early before treatment assignment is revealed.

First, to replicate among participants in the control group the same (initial) effort as in the treatment group, we asked those in the control group to upload an image of their media download settings (same as in the treatment group), but without asking them to turn off their automatic download. Presumably, this approach would screen out people unlikely to comply because of an inability to take and/or upload screenshots, but without actually changing the WhatsApp settings of those in the control group. Second, by then asking the control group to spend three days without consuming multimedia on WhatsApp, we can reliably assume that the participants in the control group would be as likely as the participants in the treatment group to agree to refrain from consuming such content, thereby increasing the internal validity of the experiment. Third, the treatment assignment (that is, the duration of multimedia deactivation) was revealed only after participants uploaded a screenshot of their WhatsApp download settings page. In this way, treatment assignment is revealed only among likely compliers, exposing participants to their assignment (three weeks or three days of deactivation) conditional on their decision to upload an image of their WhatsApp configuration.

These design choices effectively screened out participants who would be unlikely to take the treatment regardless of whether it is offered prior to treatment assignment, thus making the sample of participants across treatment groups largely comparable. We enrolled 400 participants in the treatment and 373 in the control group, resulting in respectively 70% and 65% of participants who decided to join the study after we invited them. In SI Appendix Section 3.1, we present an attrition analysis between the full sample of enrolled participants (N=773) and participants who did not join the experiment after our invitation (N=361), as well as an analysis comparing the enrolled treatment and control participants. Baseline characteristics of those who decided not to join the experiment versus those who successfully enrolled are largely balanced across fourteen sample characteristics (SI Appendix, Table 3), except for age, interest in politics, and the number of active social media accounts. We take this as an indication that

more digitally savvy participants were more likely to engage in the experiment. We control for these variables in our covariate-adjusted models. More importantly, the treatment and control groups among the enrolled are largely balanced (SI Appendix, Table 4).

Compliance Measures: We asked participants in both the treatment and control groups to upload a screenshot of their WhatsApp storage information during each of the three weeks of the experiment through a series of short Qualtrics surveys.[10] The WhatsApp storage page records the volume of bytes of media downloaded through the application. For individuals with the automatic download feature turned off (as requested of the treated participants), changes in the storage information would only occur if they purposefully clicked on multimedia they received, as that content would only be downloaded and available for viewing if they clicked on it. Users do not need to additionally save the multimedia to their devices for WhatsApp to record the action of downloading it. Importantly, the statistics on the storage page do not change even if participants decide to delete media from their phones later on, nor can the statistics be reset without leaving a record.

We use four criteria to define compliers among the treated: i) submit at least two out of the three screenshots; ii) successfully submit the third compliance image, with the information about media consumption at the end of the study; iii) have a difference in consumption in bytes smaller than the average variation among the control group; iv) not have reset their storage statistics during the weeks of the study. Importantly, while items i), ii), and iv) consider compliance as a task, item iii) considers the overall reduction in multimedia consumption in bytes to define compliers using the control group as a baseline. To retrieve the bytes information and check for other deviations from our recommendations, we manually checked all the screenshots submitted by participants. According to this criteria, we achieved a 74.8% compliance rate among the treated who completed the final survey.

[10] We provided participants with two days to complete each compliance survey. In the first compliance check, we reiterated to the control group that they could use WhatsApp normally.

Post-Treatment Survey: We followed the deactivation experiment with a comprehensive survey sent to participants on October 5, three days after the election took place. We focus our main analysis on the four sets of pre-registered outcomes: exposure and belief accuracy corresponding to a set of true stories and false rumors that circulated during the election period, polarization measures, and subjective well-being (see SI Appendix Section 2 for a full description of the various outcomes). We collected complete responses from 374 participants in the treatment group and 358 participants in the control group. Attrition following assignment, as measured in the post-treatment survey, was only 5.3%, with a slightly higher rate of 6.5% among the treatment group compared to 4% in the control group. These low rates of attrition post-treatment – especially compared to attrition prior to treatment assignment – provide evidence that our strategy to screen out *never-takers* early in the design was effective. The low levels of attrition also suggest our financial incentives were compelling for putting in the effort required to successfully participate in the experiment – participants received $30.00 (R$ 150.00) to complete the study and could qualify for three gift cards of approximately $100.00 (R$ 500.00) for completing the post-treatment survey. To mitigate concerns about bias, we compare pre-treatment demographics in SI Appendix Section 3, table 3, between those who completed and those who did not respond, and between those in treatment and those in control. We find no statistically significant imbalance between treatment and control across a range of fourteen pre-treatment variables.

Finally, because our population of interest is WhatsApp users, the validity of our design would also be at stake if the final sample were composed of participants who are otherwise not very active on WhatsApp. To alleviate these concerns, we show descriptive information about WhatsApp usage in SI Appendix Section 13, verifying that that possibility does not pose a serious threat. Specifically, Figure 14 shows that, according to pre-treatment data, over 80% of our sample uses WhatsApp for at least one hour per day. More importantly, considering exposure to politics on WhatsApp, almost 60% of the participants in the pre-treatment survey reported receiving media related to politics and elections on WhatsApp at least once every day (SI Appendix Figure 15), 70% reported using WhatsApp to consume political news at least once

a day, and 52% use WhatsApp to share/receive political information with/from their friends at least once a day (SI Appendix Figure 16). These pre-treatment statistics show that our sample's behavior mirrors the high levels of WhatsApp usage in Brazil, especially its use for political purposes (Rossini, Baptista, de Oliveira and Stromer-Galley, 2021; Newman et al., 2021).

Results

Our pre-registered analyses consist of two Intention-to-Treat (ITT) models and one Complier Average Causal Effects (CACE) model. For the ITT models, we report unadjusted OLS (ITT) and covariate-adjusted estimates (Cov-ITT)[11] of treatment effects using HC2 robust standard errors in all analyses and p-values from two-tailed t-tests. CACE models use an IV setup, instrumenting the compliance measure with an indicator for treatment assignment, and use the same set of covariates of the covariate-adjusted estimates ITT. Results adjusting for multiple hypothesis testing are presented in section 11 of the SI Appendix.

Deactivation Effects on Usage and Consumption

We first discuss the first-stage effects of the multimedia deactivation to verify that the treatment indeed reduced WhatsApp usage – particularly through multimedia content – as intended. We present results using behavioral data collected through the WhatsApp storage screenshots requested for the compliance checks. Additionally, we report results using self-reported compliance measures for accessing multimedia on WhatsApp and overall daily usage of WhatsApp as collected in the post-treatment survey. On average, participants assigned to the treatment group consumed 85.1 MB of multimedia during the three weeks of the experiment, while participants assigned to the control consumed 227 MB (two-tailed mean comparison, $t = -3.89$ and p-value < 0.01). Considering the median user as a reference, these values are 10 and 100 megabytes for treatment and control respectively (see Figure 4 in SI Appendix for the full distribution). Figure 2 presents the first stage of intention-to-treat effects on standardized

[11] Covariate-adjusted models use the following pre-registered set of pre-treatment variables: respondents' demographics (sex, age, income, race), measures of trust in mainstream media and institutions, affective polarization, ideology, interest in politics, exposure to misinformation, daily WhatsApp usage, and for which purposes participants report using WhatsApp.

coefficients for the behavioral and self-reported outcomes. The deactivation experiment reduced the consumption of multimedia on WhatsApp by 0.27 SD ($t = -3.88$, p-value < 0.01), by 0.75 SD ($t = -11.1$, p-value < 0.01) on self-reported multimedia consumption, and, unexpectedly, by 0.28 SD ($t = -3.93$, p-value < 0.01) on self-reported daily WhatsApp usage. These effects indicate that the deactivation not only effectively decreased the consumption of images, videos, and audio on WhatsApp but also led to a small reduction in overall self-reported WhatsApp usage.

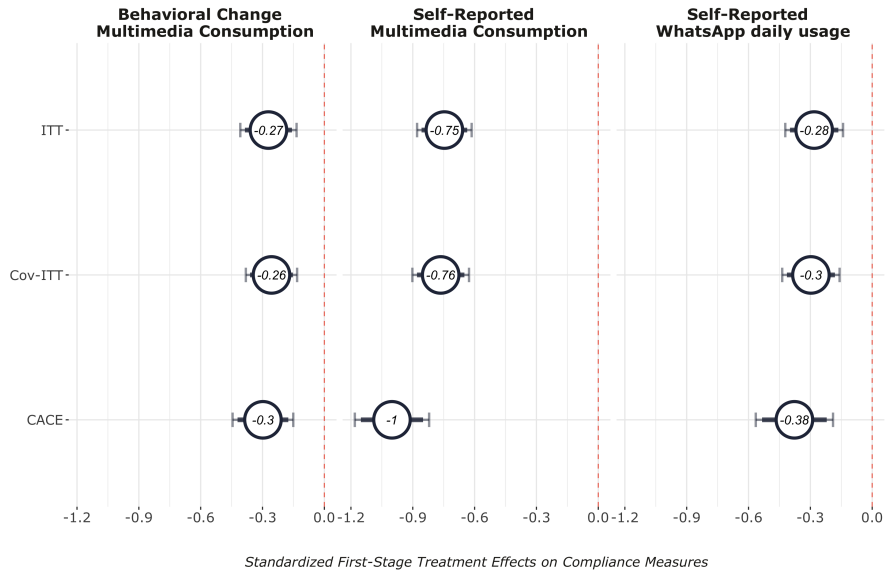


Figure 2 First stage effects of the deactivation on behavioral and self-reported WhatsApp usage.

Deactivation Effects on Exposure to Misinformation and Beliefs

To measure exposure to and belief in false rumors and true news stories, we used a set of four true news stories and four false rumors that circulated online during the period of the experiment. We randomized the appearance of these stories in the survey, and for each, we asked participants: a) to assess the accuracy of the central claim of the story and b) to indicate whether they have seen this rumor (or any closely related story) on WhatsApp in the last few weeks (see SI Appendix, Section 2, and Table 1 for the stories). The false rumors were based on news checked by four fact-checking agencies in Brazil, while the true news stories came from

mainstream news outlets (see SI Section 3 for the full description of the selection criteria for true and false stories). The eight stories are balanced across pro-left wing, pro-right wing, and non-partisan content. To measure exposure to online misinformation, we create a *Misinformation Exposure* index summing up the number of false stories which respondents reported having seen.

Figure 4 shows confirmatory evidence that the deactivation successfully reduced users' exposure to online misinformation during the pre-election weeks ($H1$). The intention-to-treat analysis shows a reduction in exposure to misinformation rumors of 0.38 SD (p-value < 0.01). Furthermore, we also find a decrease in self-reported exposure to true news stories ($d = -0.27$, p-value < 0.01). Such a result was expected: survey data indicates that Brazilians rely heavily on WhatsApp to consume news and learn about politics (Rossini, Baptista, de Oliveira and Stromer-Galley, 2021; Newman et al., 2021; Rossini, Stromer-Galley, Baptista and Veiga de Oliveira, 2021).

However, the effect sizes on reduction in exposure are considerably larger for false as opposed to true news. In relative terms, the difference in effects implies that changes in exposure to misinformation are 34% higher than exposure to true news stories. In the appendix (SI Appendix, section 12), we use ITT item-level models with clustered standard errors at the respondent level to show that larger reductions in exposure to false rumors are statistically significant at conventional levels (p-value $= 0.02$).[12] These differences are substantively significant. The larger effect size for false news shows our initial expectations were well-grounded. WhatsApp appears to be a fertile environment for the circulation of misinformation, highlighting the degree to which WhatsApp plays a more significant role in misinformation consumption relative to factual information.

Illusory Truth Effects: Does exposure affect accuracy?

So far, our results indicate the experiment worked as expected, showing causal evidence for the role of multimedia content on WhatsApp on exposure to misinformation at a larger degree than on factual information. Our next analysis focuses on the downstream effects on beliefs

[12] These results are robust to using post-treatment self-reported exposure to misleading or false information on WhatsApp during the weeks of the experiment instead of the stories task (SI Appendix, table 7).

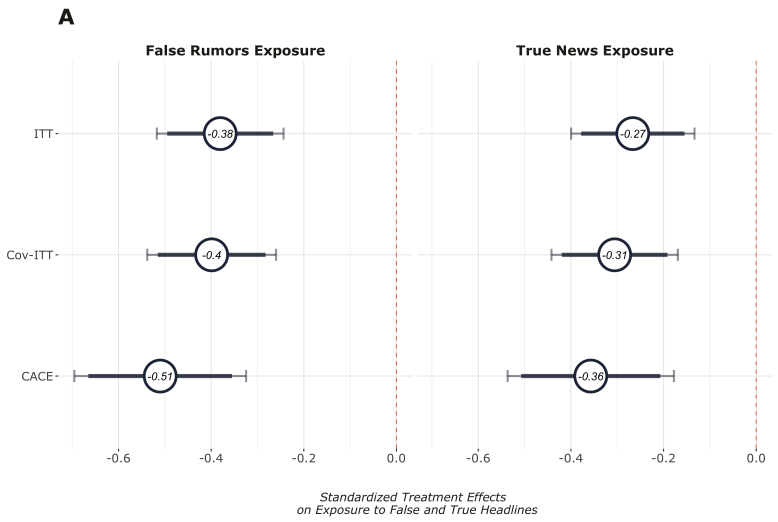


Figure 3 Treatment effects of WhatsApp Multimedia Deactivation on exposure to false rumors and true news stories. The figure presents standardized coefficients with their corresponding 95% and 90% confidence intervals based on robust standard errors.

for false and true news. To evaluate these effects, we use participants' accuracy assessments of our set of eight false rumors and true news stories. We first build two four-item indexes: *False Rumors Accuracy*, which counts the number of false items participants correctly reported as false, and *True News Accuracy*, which counts the number of true items participants correctly reported as false.

We uncover results showing a complex dynamic between exposure to misinformation and beliefs. Although the deactivation significantly reduced exposure to false and, to a lesser extent, true news, we find no statistically significant effect on belief accuracy for either true or false news. In SI Appendix Figure 8, we aggregated these measures to build a truth discernment measure, and the corresponding results are also null. We additionally investigate the presence of heterogeneity across partisan groups for improvements in *False Rumors Accuracy*, *True News Accuracy*, and recover null results (SI Appendix, Figure 9). It is important to note that previous deactivation studies also detected null effects for changes in beliefs for false news (Allcott et al., 2020). These studies only recovered statistically significant effects on mainstream news knowledge. Therefore, we consider the null effects for *True News Accuracy* to be in line with

previous studies.

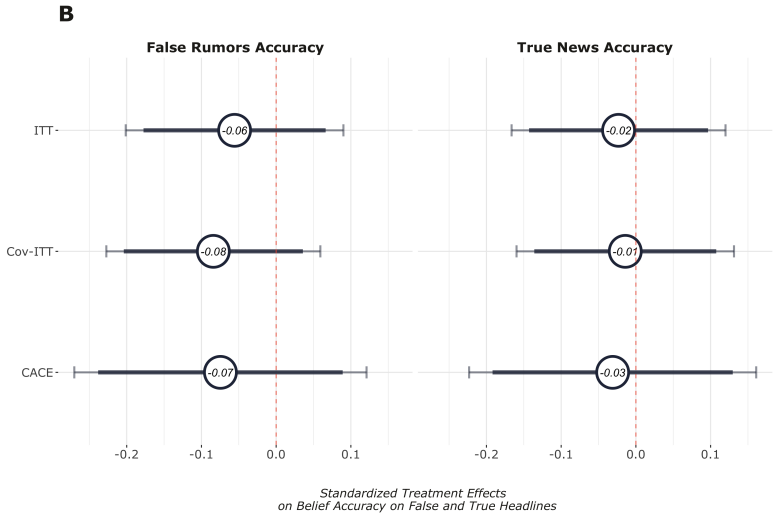


Figure 4 Treatment effects of WhatsApp Multimedia Deactivation on False Rumors Accuracy and True News Accuracy judgments. The figure presents standardized coefficients with their corresponding 95% and 90% confidence intervals based on robust standard errors.

To further unpack these null findings, we examine once more the possible theoretical mechanism that exposure to misinformation induces belief accuracy (Pennycook, Cannon and Rand, 2018). In SI Appendix Figure 5, we show that the reduction in exposure is not uniform across the four false rumors items. In addition, we show that right-wing voters reported a considerably larger reduction in exposure to misinformation than left-wing voters did (SI Appendix, Figure 6). However, the reduction in exposure is skewed, which leads to uneven exposure to ideologically congruent and incongruent misinformation. For left-wing voters, the deactivation reduced exposure to *ideologically incongruent fake news*, while for conservative voters, the deactivation decreased exposure to *congruent misinformation* (SI Appendix, Figure 7). This competitive dynamic might help us understand why a reduction in exposure does not lead to a direct improvement in belief accuracy as we had pre-registered. As descriptive research has shown, most of the supply of misinformation in Brazil aligns with the right-wing candidate Jair Bolsonaro (Avelar, 2019; Burgos, 2019; Aruguete, Calvo and Ventura, 2021; Recuero, Soares and Vinhas, 2021); thus, instead of a uniformly distributed effect, the deactivation pushes some subjects

away from false news they are likely to reject and others from false news they are likely to believe, which may, in turn, affect how exposure translates to accuracy judgments.

In our pre-analysis plan, we proposed to examine heterogeneous effects for accuracy improvements conditional on age, digital literacy, and the use of WhatsApp for politics. Our expectation was to detect larger improvements in belief accuracy among older subjects, subjects with lower levels of digital literacy, and subjects who frequently reported receiving political content on WhatsApp. We find null effects related to age and digital literacy but observe that improvements in identifying false rumors vary based on how frequently subjects reported receiving political information on WhatsApp (as measured in the pre-treatment survey) (See SI Appendix, section 8). For participants who rarely receive political information, their accuracy judgments for false rumors became worse post-treatment, while for participants who reported receiving politics on WhatsApp multiple times a day, the deactivation improved their capacity to identify false rumors, as we expected in $H2a$ (SI Appendix, Table 10, and figure 5. We consider this interesting heterogeneity only as suggestive evidence that reducing exposure to misinformation on WhatsApp leads to larger improvements in belief accuracy among participants who rely heavily on WhatsApp for politics, and an informational decay among participants with lower baseline exposure to political information on WhatsApp. We return to this result in the discussion section.

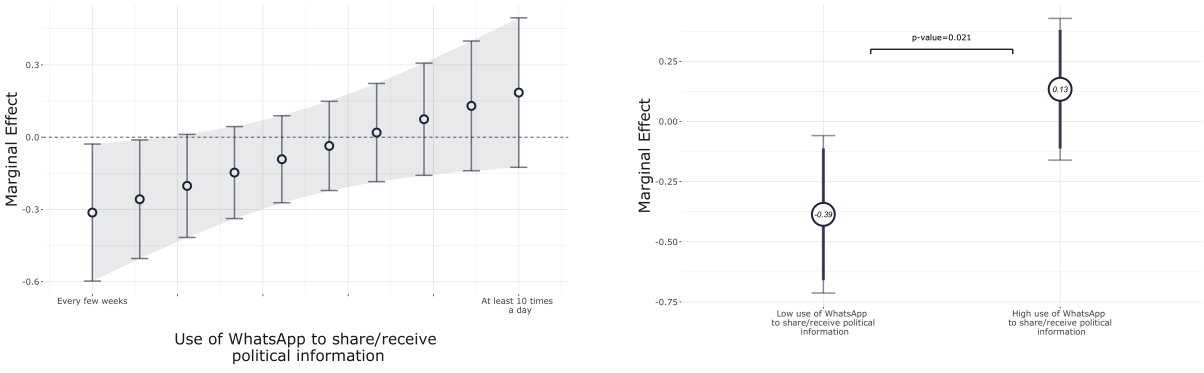


Figure 5 Marginal Treatment effects of WhatsApp Multimedia Deactivation on Belief Accuracy conditional on using WhatsApp to receive and share information about politics. The left plot shows conditional effects using a continuous scale of the moderator values. The right plot uses the difference between low-dosage users (first-tercile of the continuous scale) and heavy-dosage users (third-tercile of the continuous scale)

Deactivation Effects on Polarization

We turn now to treatment effects on polarization. We pre-registered hypotheses focusing on a composite index combining four distinct polarization measures: absolute differences in feeling thermometer responses about two front-running presidential candidates (*affective polarization*), willingness to engage in five distinct social activities with outgroup voters (*social polarization*), misperceptions of ideological polarization across the two main candidates (*false polarization*), and polarization across five distinct policy issues (*issue polarization*).

We present results that suggest adopting a minimalist view of the short-term effects of exposure to misinformation via WhatsApp on polarization. Using our pre-registered aggregate polarization index as the primary outcome, we present treatment effects in Figure 6. We cannot reject the null hypothesis of no effect of the deactivation on outgroup political polarization. As a way to assess the robustness of our findings, we also examine results for each of the four measures of polarization that form our aggregate index (SI Appendix, Figure 10). Most results across the four distinct measures of polarization are null, as in the polarization index, except for a modest reduction in the social polarization outcome ($d = -0.13$, unadjusted p-value $= 0.08$). Even though we report null effects of the deactivation on polarization, we believe that the salience of social polarization among the other polarization measures deserves further discussion in light of how communication networks are built on WhatsApp.

Unlike other social media applications such as Facebook and Twitter, WhatsApp does not have an organized news feed in which users receive/are exposed to information posted by their friends, news organizations, or political authorities. WhatsApp communication networks rely mainly on users' contacts; information exchanges are more personal and flow through one-to-one or group chat communications. As discussed previously, social polarization measures participants' willingness to engage in a list of four different social activities with individuals who voted for a different presidential candidate. This is the only one of our four polarization measures that captures outgroup attitudes about other voters rather than about elites, candidates, or policy issues, and interestingly the measure with the largest negative effects. As such, we encourage future studies to focus further on social polarization as a primary outcome when

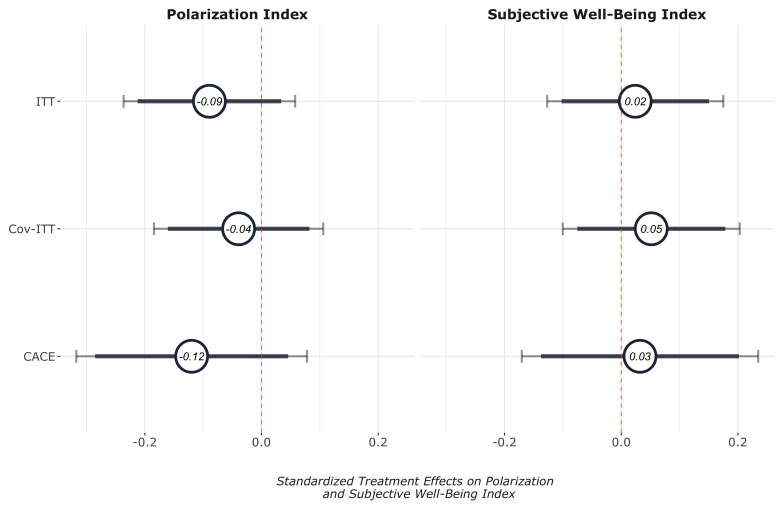


Figure 6 Treatment effects of WhatsApp Multimedia Deactivation on Polarization and Subjective Well-Being. The figure presents standardized coefficients with their corresponding 95% and 90% confidence intervals based on robust standard errors.

investigating the effects of social media messaging apps like WhatsApp.

We also highlight that our null, pre-registered findings for polarization are consistent with recent experiments on the effects of news exposure on attitudes and behavior (Guess et al., 2021; Aslett et al., 2022). More importantly, results from similar deactivation studies are rather mixed. For example, Allcott et al. (2020) find a decrease in polarization on average after their month-long Facebook deactivation study. Meanwhile, Asimovic et al. (2021) find in their deactivation study in Bosnia and Herzegovina that the effects run in the opposite direction, and in a deactivation study in Cyprus, Asimovic, Nagler and Tucker (2023) find no effect of reducing social media usage on polarization. Our null results add more evidence to this mix of results and reinforce a minimalist view of the direct effects of social media on polarization. Indeed, null results on polarization were also recovered from a recent set of well-powered studies on Facebook and Instagram during the 2020 US presidential (Guess, Stroud and Tucker, 2023; Guess et al., 2023; Nyhan et al., 2023).

Deactivation Effects on Subjective Well-Being

We pre-registered secondary hypotheses focusing on subjective well-being as an outcome. We asked participants about five different measures: happiness, satisfaction, social isolation, anxiety, and depression. We standardize all measures and use the sum of their z-scores to build a composite index. We recover null effects for the index of subjective well-being. No heterogeneity is detected conditional on age, partisan leaning, digital literacy, and the use of WhatsApp for politics. To ensure results are robust, we also present results for all five subjective well-being items in the SI Appendix Figure 11. We obtain null results across the board.

When put in perspective with results from previous studies, these null effects help inform the pathways through which social media usage might affect subjective well-being. Social media applications can affect subjective well-being through changes in information consumption or through changes in daily habits. Our experiment only manipulated the medium through which subjects received information, reducing their exposure to multimedia on WhatsApp (and their subsequent exposure to misinformation). Full deactivation studies performed on other social media platforms (Allcott et al., 2020; Asimovic et al., 2021; Vanman, Baker and Tobin, 2018) manipulate both the information diets and daily usage of social media for participants. Our null results suggest that the main driver of recently reported psychological effects of social media usage might not be the former, but rather how social media usage shapes users' daily habits.

Conclusion

There is widespread concern among researchers, journalists, and politicians that social media plays a crucial role in facilitating the spread of online misinformation, and in turn, shaping citizen's attitudes. In many countries in the Global South, 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. Inspired by previous work on identifying the political effects of Facebook usage, we implemented a novel deactivation field experiment on WhatsApp in which we incentivized users in Brazil to not access any image, audio, or video received on WhatsApp

during the three weeks leading up to their general election on October 2nd, 2022.

We provide casual evidence that this multimedia deactivation reduced exposure to online misinformation. Subjects in the treatment group reported a substantive reduction in exposure to four distinct false rumors that fact-checking agencies flagged during the election campaign. These effects are larger than the observed reduction in exposure to true news stories. Beyond changes in exposure, we do not capture treatment effects for improvements in accuracy judgments for the same false and true stories, reduction in political polarization, or improvements in subjective well-being. Similar null effects have emerged from field experiments manipulating, for example, exposure to online partisan news and untrustworthy websites (Guess et al., 2021, 2020), news credibility labels on online news sources (Aslett et al., 2022), and exposure to mainstream news (Wojcieszak et al., 2022). More importantly, previous deactivation studies have shown that when taken out from social media, users tend to become less knowledgeable only about true news, but no effects have been detected for accuracy judgments of misinformation (Allcott et al., 2020) – similar to the results in our experiment. Taken together, these results indicate more research is needed to document the varied short-term effects of consuming online misinformation (Guess et al., 2020). Furthermore, these findings are consistent with the minimal-effects tradition of news exposure on attitudes and challenge arguments connecting in a mechanical manner exposure to news and misinformation rumors to belief formation.

At the same time, we present suggestive evidence of heterogeneous treatment effects conditional on self-reported usage of using WhatsApp as a tool to receive political news. Those who report receiving political content on WhatsApp multiple times a day improve their capacity to identify false rumors, while those who report rarely receiving political news via WhatsApp become significantly worse at the task of identifying misinformation in our post-treatment survey. Similar heterogeneous effects have been detected in prior studies (Aslett et al., 2022) and speak to the importance of investing more in research that oversamples heavy online information consumers.

Now, despite the set of null results produced by our experiment, we warn that it would be a mistake to conclude that WhatsApp plays no role in politics. Our research is limited in scope

to detect the direct effects of imposing constraints on consuming misinformation via WhatsApp on political attitudes on the verge of a crucial political event. This limited scope does not rule out the use of WhatsApp as a coordination and mobilization tool, particularly the use of the platform by harmful actors to mobilize and organize supporters for offline activities, such as the events of January 8th, 2023 in Brazil, when thousands of supporters of former President Jair Bolsonaro stormed key government buildings to challenge Brazil's democratic institutions.

Furthermore, our choice of running the experiment in the weeks before a critical election brings both strengths and weaknesses. Our design is well-suited to capture a context in which the effects of social media are substantively important. However, election periods also provide voters with a high-information environment, in which social media and news are flooded with political information (including misinformation), as well as a high-choice media environment. For example, participants in the treatment condition reported watching more television during the experimental period (see Table 12, SI Appendix), though no substitution effects appeared for other social media platforms. Additionally, election dynamics can increase citizens' reliance on motivated reasoning and the partisan nature of online rumors. Theoretically, motivated reasoning can act as a potential boundary condition of our primary cognitive mechanism, that is, the "illusory truth effect" (Pennycook, Cannon and Rand, 2018), and consequentially mitigate the effects of our deactivation treatment. Replications of this design in non-election periods would be a valuable next step in advancing our understanding of the causal mechanisms linking WhatsApp usage and beliefs in misinformation and polarization.

Our final caveat relates to the statistical power of our design.[13] Our study is well-powered to detect small effect sizes (80% of power for Cohen's d >0.2 SD), similar to the effects detected in similar studies. However, we must warn the readers to consider our null effects with caution in light of the presence of even smaller effects on the population, which our design is not powered to detect. We encourage future studies to replicate our design, as has happened in replications of previous Facebook deactivation designs (Asimovic, Nagler and Tucker, 2023; Allcott et al., 2024), to test the robustness of our null results.

[13] We direct the reader to the appendix to see our pre-registered power analysis

Despite these caveats, these are important scientific and policy takeaways from our study. First, we come to similar conclusions on a different platform in a different country as recent studies of the impacts of Facebook and Instagram during the U.S. 2020 election (Guess et al., 2023; Guess, Stroud and Tucker, 2023; Nyhan et al., 2023) have: simple adjustments to how users engage with social media platforms are not sufficient on their own to impact important political attitudes. In the same vein, similar to Guess, Stroud and Tucker (2023), we do, however, find that a fairly simple tweak – in our case, increased friction to accessing videos and images – does reduce exposure to misinformation online and affects accuracy beliefs for heavy WhatsApp users. The caveat, though, is that like Guess, Stroud and Tucker (2023) found on Facebook, we also find that this reduction in exposure to misinformation is accompanied by a reduction in exposure to political news generally. Furthermore, as in other similar studies (Aslett et al. 2022), our findings offer yet more evidence that even when interventions do not have an impact on the population writ large, there is a possibility that they may impact a sample of interest at the tails of the distribution.

Finally, our study shows the continued importance of moving beyond what we know about the impact of social media usage on politics in the United States (Tucker et al., 2018). As the vast majority of social media users reside outside of the United States, we must continue subjecting what we think we know about social media to different theoretical approaches and empirical tests, spanning not only various geographic contexts, but also different platform types that enjoy tremendous popularity elsewhere. Our research on WhatsApp use in Brazil is an important step in that direction, but much more remains to be done in this regard.

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.

Aruguete, Natalia, Ernesto Calvo and Tiago Ventura. 2021. "News sharing, gatekeeping, and polarization: A study of the# Bolsonaro Election." *Digital journalism* 9(1):1–23.

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*a*. "Researching and Countering Misinformation in the Global South." *Current Opinion in Psychology* p. 101733.

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

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.

Banaji, Shakuntala, Ramnath Bhat, Anushi Agarwal, Nihal Passanha and Mukti Sadhana Pravin. 2019. "WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India." *Working Paper* .

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.

Bennett, W Lance and Shanto Iyengar. 2008. "A new era of minimal effects? The changing foundations of political communication." *Journal of communication* 58(4):707–731.

Bode, Leticia. 2016. "Political news in the news feed: Learning politics from social media." *Mass communication and society* 19(1):24–48.

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

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.

Dechêne, Alice, Christoph Stahl, Jochim Hansen and Michaela Wänke. 2010. "The truth about the truth: A meta-analytic review of the truth effect." *Personality and Social Psychology Review* 14(2):238–257.

Del Vicario, Michela, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley and Walter Quattrociocchi. 2016. "The spreading of misinformation online." *Proceedings of the National Academy of Sciences* 113(3):554–559.

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.

Fazio, Lisa K, Nadia M Brashier, B Keith Payne and Elizabeth J Marsh. 2015. "Knowledge does not protect against illusory truth." *Journal of Experimental Psychology: General* 144(5):993.

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

Freitas Melo, Philipe de, 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.

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*.

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.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson and David Lazer. 2019. "Fake news on Twitter during the 2016 US presidential election." *Science* 363(6425):374–378.

Guess, Andrew M, Dominique Lockett, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan and Jason Reifler. 2020. ""Fake news" may have limited effects beyond increasing beliefs in false claims." *Harvard Misinformation Review* .

Guess, Andrew M., Neil Malhotra ...., 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.

Guess, Andrew M., Neil Malhotra .... Natalie Jomini Stroud and Joshua A. Tucker. 2023. "Reshares on social media amplify political news but do not detectably affect beliefs or opinions." *Science* 381(6656):404–408.

Guess, Andrew M, Pablo Barberá, Simon Munzert and JungHwan Yang. 2021. "The consequences of online partisan media." *Proceedings of the National Academy of Sciences* 118(14):e2013464118.

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.

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.

Lazarsfeld, Paul F, Bernard Berelson and Hazel Gaudet. 1968. *The people's choice: How the voter makes up his mind in a presidential campaign*. Columbia University Press.

Lorenz-Spreen, Philipp, Lisa Oswald, Stephan Lewandowsky and Ralph Hertwig. 2022. "A systematic review of worldwide causal and correlational evidence on digital media and democracy." *Nature human behaviour* pp. 1–28.

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.

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

Mitchelstein, Eugenia and Pablo J. Boczkowski. 2021. "What a Special Issue on Latin America Teaches Us about Some Key Limitations in the Field of Digital Journalism." *Digital Journalism* 9(2):130–135.

Nanz, Andreas and Jörg Matthes. 2022. "Democratic Consequences of Incidental Exposure to Political Information: A Meta-Analysis." *Journal of Communication* 72(3):345–373.

Newman, Nic, Richard Fletcher, Anne Schulz, Simge Andi, Craig T Robertson and Rasmus Kleis Nielsen. 2021. "Reuters Institute digital news report 2021." *Reuters Institute for the study of Journalism* .

Nyhan, Brendan, Jaime Settle, Natalie Jomini Thorson, Emily .... Stroud and Joshua A. Tucker. 2023. "Like-minded sources on Facebook are prevalent but not polarizing." *Nature* .

Park, Chang Sup and Homero Gil de Zúñiga. 2020. "Learning about Politics from Mass Media and Social Media: Moderating Roles of Press Freedom and Public Service Broadcasting in 11 Countries." *International Journal of Public Opinion Research* 33(2):315–335.

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.

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, 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.

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.

Suhay, Elizabeth, Emily Bello-Pardo and Brianna Maurer. 2018. "The polarizing effects of online partisan criticism: Evidence from two experiments." *The International Journal of Press/Politics* 23(1):95–115.

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

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)* .

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.

Vosoughi, Soroush, Deb Roy and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359(6380):1146–1151.

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.

Misinformation Exposure Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil

Supporting Information Files (SIF)

Contents

1 Survey Materials

We present here some of the most important pieces of our survey design. All survey materials are presented in their full formats at our pre-registration. Our research was approved by the [REDACTED] Institutional Review Board (protocol IRB-FY2022-6727X). Subsection 1.1 presents the Recruitment Ad displayed on Facebook. Upon clicking on the ad, participants were taken from Facebook to the pre-treatment survey on Qualtrics. The Ad was active between September 8 through September 14. Subsection 1.2 presents the screenshots requested as part of the multimedia deactivation. Subsection 1.3 presents the images requested for the compliance checks. All materials were presented to participants in Portuguese. We communicate with participants through email and WhatsApp, while using Qualtrics for all the survey questionnaires, including the collection of images for the compliance checks. Original materials in Portuguese are available upon request.

1.1 Preview of the Recruitment

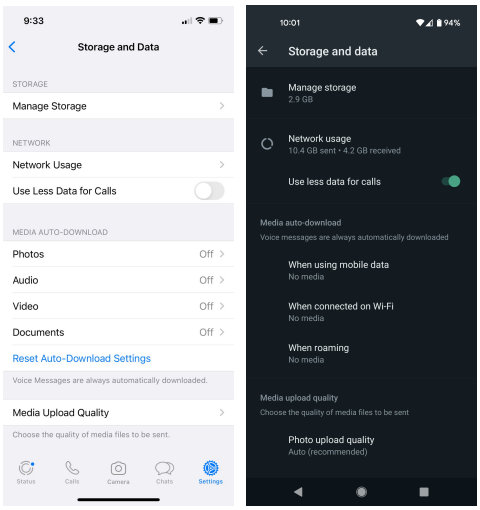
Figure 1 Facebook Advertisement Used for Recruitment

[REDACTED]

Notes: We started recruitment on September 8 and finished on September 14

1.2 Deactivation Insturctions

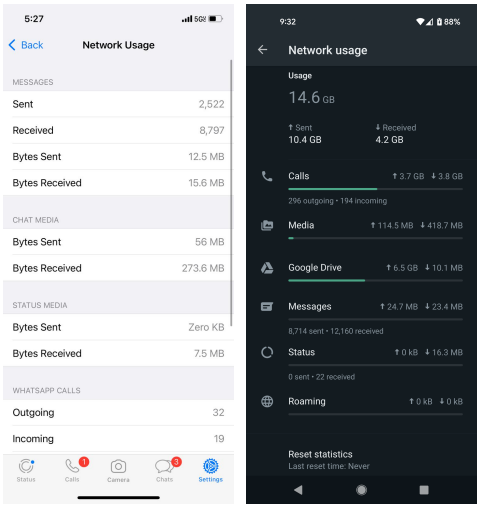
Figure 2 Images for Treatment Assignment Survey: Deactivation of automatic media download on WhatsApp. To facilitate compliance among the treated participants, we requested the treatment group to disable their automatic download of media on WhatsApp, and upload a screenshot of their apps. The left figure comes from an iPhone device and the right figure is from an Android device.



1.3 Compliance Checks

To measure compliance, we asked respondents to upload a screenshot of the storage information of their mobile WhatsApp app. We use a short Qualtrics survey to collect these images, with complete instructions about how to access these images. We present examples of the screenshots below.

Figure 3 Compliance ChecksThe left figure comes from an iPhone device, and the right figure is from an Android device.



2 Outcomes

To study the effect of the multimedia deactivation on belief accuracy and exposure to misinformation, we conducted a task using a set of true news stories and false rumors that circulated online during the period of the experiment. Respondents repeat this task eight times. For each headline, we asked participants: a) "What is your assessment of the central claim in the article?" to which respondents could choose from three responses: (1) True (2) Misleading/False (3) Could Not Determine; and "In the past 30 days, do you remember seeing or receiving this headline on WhatsApp?" to which participants could choose from: (1) yes, (2) no, (3) I don't remember.

We propose a transparent strategy to select the stories to mitigate bias from the selection of misinformation stories included in the final survey. During the weeks of the experiment, we scraped the four of the main fact-checking pages in Brazil (*Comprova, Aos Fatos, Boatos, Lupa*) to collect stories checked as false or misleading that circulated during the weeks of the experiment. We classify the stories according to their topic, political leaning, how many fact-checking agencies checked the story, and if it circulated on WhatsApp. Table 1 presents the full wording for each of the eight headlines. We included stories that most agencies checked and selected diverse stories considering their topic and political leaning. Due to the end-to-end encryption of WhatsApp, there is no straightforward way to capture the most salient stories circulating in the app. We believe our strategy of focusing on stories checked by multiple fact-checkers works as an effective proxy to capture most salience rumors circulating online.

Out of the four misinformation stories, we used two stories that are pro-attitudinal to Bolsonaro voters, one pro-attitudinal to Lula's voters, and one non-partisan story. This imbalance is a consequence of the fact that right-wing misinformation prevails in Brazil (Burgos 2019; Recuero, Soares and Vinhas 2021; Aruguete, Calvo and Ventura 2021). Analyzing all stories scraped from the four

above-mentioned fact-checking organizations, we found a similar pattern for the 2022 election. We kept the same partisan distributions for the TRUE headlines in order to avoid signaling to participants a larger share of FALSE/TRUE headlines conditional on their partisan direction. The true stories were collected from mainstream media news outlets.

We added four different variables to measure political polarization: affective polarization, false polarization, social polarization, and issue polarization. We consider all these measures to capture different dimensions of political polarization, and, based on this assumption, we pre-registered $H3$ using a *polarization index* built using z-scores across all these variables as our primary outcome of interest. For the secondary hypotheses, we pre-registered the hypothesis using an index of subjective well-being combining individual responses to our outcomes. We use five different outcomes for subjective well-being (Happy, Depressed, Anxious, Isolated from my family and friends, Satisfied with my life). Table 2 fully describes the outcomes separately.

Table 1 True news stories and false rumors selected to measure exposure to misinformation and belief accuracy. The order of the items was fully randomized.

Veracity	Stories in Portuguese	Stories in English	Direction
False Item 1	Diversas pesquisas de intenção de voto mostravam Bolsonaro liderando a corrida eleitoral, com possibilidades de ganhar a eleição em primeiro turno	Several voting intentions polls showed Bolsonaro leading the electoral race, with possibilities of winning the election in the first round	Pro-Bolsonaro
False Item 2	Somente votos completos são computados pela justiça eleitoral. Caso o eleitor vote só para presidente, e votar em branco nos outros, o voto é tido como voto parcial, e será anulado.	Only complete votes are counted by the electoral justice. If the voter only votes for the President, and votes blank for all the other races, the vote is considered a partial vote, and will be annulled	Non-Partisan
False Item 3	Antes de iniciar a votação do primeiro turno, a Polícia Federal identificou urnas eletrônicas com votos já registrados em Brasília	Before starting to vote on the first turn, the Federal Police identified electronic ballot boxes with votes already registered in Brasilia	Pro Bolsonaro
False Item 4	Jair Bolsonaro, atual presidente e candidato à reeleição, foi aposentado do exército aos 33 anos por com diagnóstico de problemas psicológicos.	Jair Bolsonaro, current president and candidate for re-election, was retired from the army at the age of 33 due to a diagnosis of psychological problems	Pro Lula
True Item 1	Rússia fez nova mobilização para Guerra na Ucrânia. Nos últimos dias, o Presidente Vladimir Putin anexou à Russia territórios ocupados na Ucrânia	The Russian forces made new mobilization for War in Ukraine. In recent days, President Vladimir Putin has annexed occupied territories in Ukraine to Russia.	Non-Partisan
True Item 2	O carro e a casa da ex-mulher do presidente Jair Bolsonaro, a candidata a deputada distrital Ana Cristina Valle foram depredados e pichados nesta última semana de eleição	Last week, the car and house of President Jair Bolsonaro's former-wife, and state-level legislative candidate for Ana Cristina Valle	Pro-Bolsonaro
True Item 3	Após reduções consecutivas, o preço atual da gasolina no Brasil está abaixo da média mundial	After consecutive reductions, the current oil price in Brazil is below the global average value	Pro-Bolsonaro
True Item 4	O Senado dos Estados Unidos aprovou uma resolução proposta pelo senador democrata Bernie Sanders em defesa da democracia no Brasil	The United States Senate approved a resolution proposed by Democratic Senator Bernie Sanders in defense of democracy in Brazil	Pro-Lula

Table 2 Pre-Registered Dependent Variables

Variable	Description	Scale
Misinformation Outcomes		
q_false_news_acc	Respondent's assessment of four short stories (headlines) retrieved from news checked as False by the most popular fact-checking agencies in Brazil during the time of the experiment	False, True, Unsure
q_false_news_exposure	Respondent's exposure to four short stories (headlines) retrieved from news checked as False by the most popular fact-checking agencies in Brazil during the time of the experiment	Yes, No, I don't remember
q_true_news_acc	Respondent's assessment of four short stories (headlines) retrieved from mainstream news media during the time of the experiment	False, True, Unsure
q_true_news_exposure	Respondent's exposure to four short stories (headlines) retrieved from mainstream news media during the time of the experiment	Yes, No, I don't remember
False Rumors Accuracy	Sum of false news respondent identify as false	0-4
True News Accuracy	Sum of true news respondent identify as true	0-4
False Rumors Exposure	sum of misinformation stories respondent was exposed to	0-4
Polarization Outcomes		
Polarization Index	Sum of the z-scores for four polarization outcomes	-1 - 1
Affective Polarization	Difference in the feeling thermometers between the two main presidential candidates	0-100
False Polarization	Difference in the ideological scale (q_ideology) between the two main presidential candidates	0-10
q_social_outgroup	Number of activities respondents is willing to participate with an outgroup partisan	0-6
q_issue_polarization	Agreement across five different policy issues	1-7 issue agreement
Subjective Well-Being		
Subjective Well-Being Index	Sum of the z-scores for four composite items (Items: Happy, Depressed, Anxious, Isolated from my family and friends, Satisfied with my life)	-3 - 3

3 Attrition Analysis

We invited 1,135 participants to take part in the deactivation experiment. Out of the individuals invited, 773 participants were enrolled in the experiment and collected 732 complete responses in our post-treatment survey after the three weeks deactivation period. For our study, we define two different sources of attrition. First, we define it as attrition-among-the-invited those participants invited to the experiment, but that failed to join our study. Second, we define it as attrition-among-enrolled those participants that entered the experiment, had their treatment assignment revealed, but failed to complete the post-treatment survey.

As in any experimental study with multiple waves, attrition might be a source of bias. Attrition may introduce bias if the treatment and control group attrit not at random and if covariates correlated with the missingness are systematically related to the outcomes of interest. As usual in these cases, we analyze the wide range of baseline characteristics of users for the two potential sources of differential attrition. Furthermore, we use the same baseline covariates to conduct randomization inference analysis predicting the treatment assignment both among enrolled participants and among complete participants.

3.1 Missing Among Invited

To analyze differential attrition among invited, we first present two tables. Table 3 compares baseline characteristics across 14 pre-treatment variables for participants who were invited to the experiment, but dropped out in the enrollment survey, and participants who successfully entered the experiment. There are only three out of fourteen statistically significant differences. Participants who entered the experiment were slightly younger, more interested in politics, and more active on social measured by the number of active accounts participants report to have on six different social media platforms. This result may indicate participants with high digital literacy were slightly more likely to join the study. However, when looking at differential attrition, Table 4 finds no statistically significant differences between enrolled participants among the treatment and control groups. The overall balance across both groups alleviates concerns of differential attrition causing bias in the treatment effects.

3.2 Missing Among Enrolled

We conduct similar analyses from the previous section but now we focus solely on attrition-among-enrolled. Table 5 compares baseline characteristics across the same fourteen pre-treatment variables for participants who successfully started the experiment, but dropped out before the post-treatment survey. Out of fourteen variables, only gender is barely significant, indicating a slightly higher share of women completing the study. When looking at differential attrition, Table 6 finds no statistically significant differences between enrolled participants among the treatment and control groups. The overall balance across both groups alleviates concerns of differential attrition causing bias in the treatment effects.

Table 3 Balance table comparing baseline characteristics for attrition-among-the invited. The table compares participants who dropped out of the enrollment survey and participants who successfully completed the enrollment in the experiment. We conduct a t-test across the groups in each baseline variable and report its two-tailed p-value against the null hypothesis that users in both groups are not different from each other.

	Dropout Treatment Instructions (N=358)		Enrolled in Experiment (N=773)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p.value
Age	2.81	1.24	2.28	0.93	-0.54	0.00
Gender	1.20	0.40	1.16	0.37	-0.04	0.12
Education	4.57	1.08	4.63	1.01	0.06	0.39
Interest in Politics	2.71	0.74	2.82	0.75	0.12	0.01
Ideology Self	5.60	3.49	5.33	3.21	-0.27	0.22
Trust elections	3.83	1.31	3.96	1.23	0.13	0.10
Polarization Candidates	6.62	3.38	6.74	3.46	0.12	0.59
WP:Daily time	6.77	1.40	6.85	1.32	0.08	0.36
WP:Frequency Politics	4.66	2.13	4.42	2.12	-0.24	0.09
WP:Frequency News	3.26	1.96	3.31	1.98	0.05	0.69
WP:Frequency Family	1.93	1.03	1.87	0.91	-0.06	0.35
WP:Frequency Images about Politics	3.73	2.01	3.70	1.93	-0.04	0.76
N Active Social Media Apps	3.47	1.37	3.94	1.29	0.47	0.00
Duration Pre-treatment	13.48	60.27	18.64	210.39	5.15	0.53

Table 4 Balance table comparing baseline characteristics for treatment groups among participants enrolled in the experiment. We conduct a t-test across the groups in each baseline variable and report its two-tailed p-value against the null hypothesis that users in both groups are not statistically different from each other.

	Control Enrolled (N=373)		Treatment Enrolled(N=400)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p.value
Age	2.28	0.93	2.27	0.94	-0.01	0.83
Gender	1.15	0.36	1.17	0.38	0.02	0.40
Education	4.61	1.02	4.64	1.01	0.03	0.69
Interest in Politics	2.84	0.75	2.81	0.76	-0.03	0.56
Ideology Self	5.29	3.17	5.36	3.25	0.07	0.76
Trust elections	3.94	1.24	3.98	1.22	0.04	0.62
Polarization Candidates	6.70	3.51	6.77	3.43	0.07	0.79
WP:Daily time	6.81	1.34	6.89	1.30	0.08	0.40
WP:Frequency Politics	4.48	2.13	4.37	2.11	-0.11	0.50
WP:Frequency News	3.32	2.04	3.31	1.93	-0.02	0.90
WP:Frequency Family	1.85	0.92	1.89	0.89	0.04	0.50
WP:Frequency Images about Politics	3.75	1.97	3.65	1.88	-0.10	0.49
N Active Social Media Apps	3.98	1.28	3.90	1.31	-0.08	0.37
Duration Pre-treatment	29.55	302.09	8.46	18.59	-21.09	0.18

Table 5 Balance table comparing baseline characteristics for attrition-among-the-enrolled. The table compares participants who dropped out in the post-treatment survey and participants who successfully completed the experiment. We conduct a t-test across the groups in each baseline variable and report its two-tailed p-value against the null hypothesis that users in both groups are not different from each other.

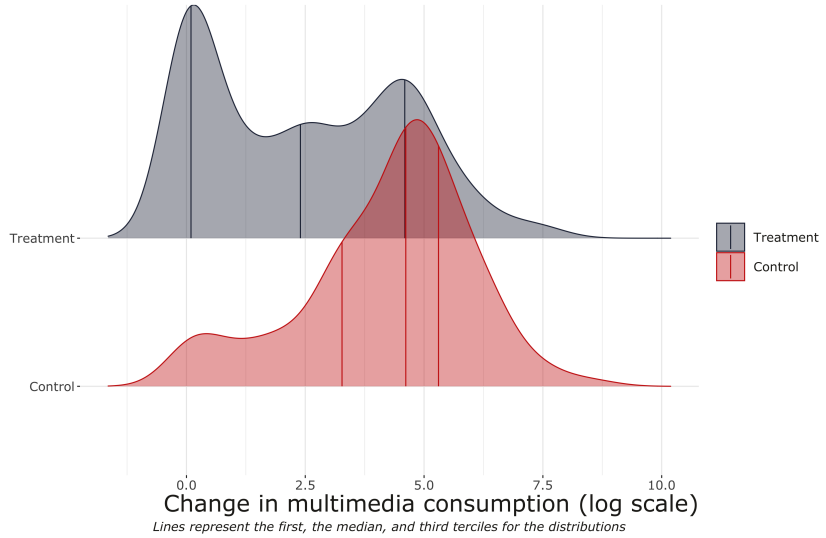
	Completes (N=732)		Dropout in Post-Treatment (N=41)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p.value
Age	2.26	0.93	2.51	0.98	0.25	0.12
Gender	1.17	0.37	1.07	0.26	-0.09	0.04
Education	4.62	1.00	4.68	1.29	0.06	0.78
Interest in Politics	2.82	0.75	2.90	0.77	0.08	0.50
Ideology Self	5.33	3.21	5.27	3.29	-0.07	0.90
Trust elections	3.95	1.22	4.12	1.42	0.17	0.46
Polarization Candidates	6.73	3.48	6.95	3.09	0.23	0.65
WP:Daily time	6.85	1.32	6.90	1.32	0.05	0.80
WP:Frequency Politics	4.44	2.12	4.21	2.12	-0.23	0.53
WP:Frequency News	3.33	1.98	3.10	2.00	-0.23	0.49
WP:Frequency Family	1.87	0.91	2.00	0.81	0.13	0.31
WP:Frequency Images about Politics	3.72	1.94	3.34	1.76	-0.37	0.19
N Active Social Media Apps	3.95	1.28	3.59	1.41	-0.37	0.11
Duration Pre-treatment	19.34	216.18	6.01	3.01	-13.34	0.10

Table 6 Balance table comparing baseline characteristics for treatment groups among participants who completed the experiment. We conduct a t-test across the groups in each baseline variable and report its two-tailed p-value against the null hypothesis that users in both groups are not statistically different from each other.

	Control Among Completes (N=358)		Treatment Among Completes (N=374)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p.value
Age	2.27	0.92	2.26	0.94	-0.01	0.90
Gender	1.15	0.36	1.18	0.38	0.03	0.35
Education	4.62	1.00	4.63	1.00	0.02	0.82
Interest in Politics	2.84	0.74	2.80	0.76	-0.03	0.55
Ideology Self	5.27	3.17	5.40	3.25	0.12	0.60
Trust elections	3.94	1.23	3.97	1.21	0.03	0.77
Polarization Candidates	6.70	3.54	6.75	3.44	0.04	0.87
WP:Daily time	6.81	1.34	6.88	1.29	0.07	0.48
WP:Frequency Politics	4.49	2.14	4.39	2.11	-0.10	0.53
WP:Frequency News	3.32	2.04	3.33	1.93	0.01	0.96
WP:Frequency Family	1.84	0.94	1.89	0.89	0.05	0.48
WP:Frequency Images about Politics	3.75	1.97	3.69	1.90	-0.06	0.68
N Active Social Media Apps	4.01	1.25	3.91	1.32	-0.10	0.30
Duration Pre-treatment	30.52	308.33	8.64	19.19	-21.88	0.18

4 Additional Results: Compliance

Figure 4 Full Distribution for the difference Megabytes of behavioral consumption data of multimedia on WhatsApp. The densities plot the differences across treatment and control participants between the first compliance check and the compliance check embedded in the post-treatment survey. Results are presented in the log scale. We remove from the graph participants that failed to submit the compliance or presented a negative variation in multi-media consumption. Negative variations might be a result of participants resetting their statistics or exogenous updates on participants' mobile apps



5 Additional Results: Exposure and Belief Accuracy

Figure 5 Treatment effects of WhatsApp Multimedia Deactivation split by every false headline. The figure presents standardized coefficients with their corresponding 95% and 90% confidence intervals based on robust standard errors.

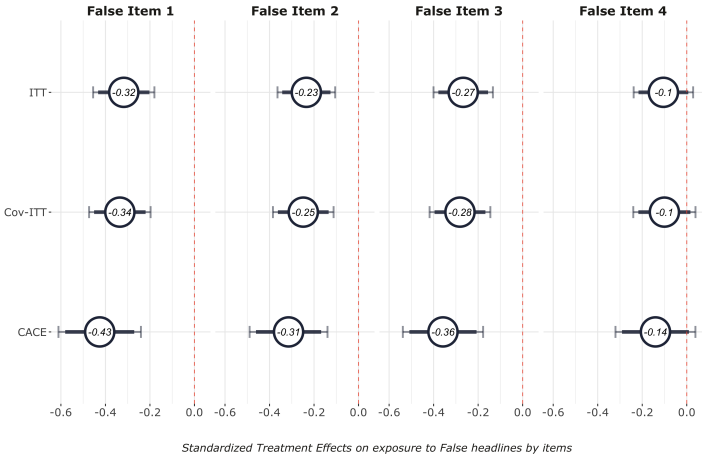


Figure 6 Treatment effects of WhatsApp Multimedia Deactivation on exposure to false information conditional on voting preferences. The models use the *False Rumors Exposure* index as an outcome. The figure presents standardized coefficients with their corresponding 95% and 90% confidence intervals based on robust standard errors.

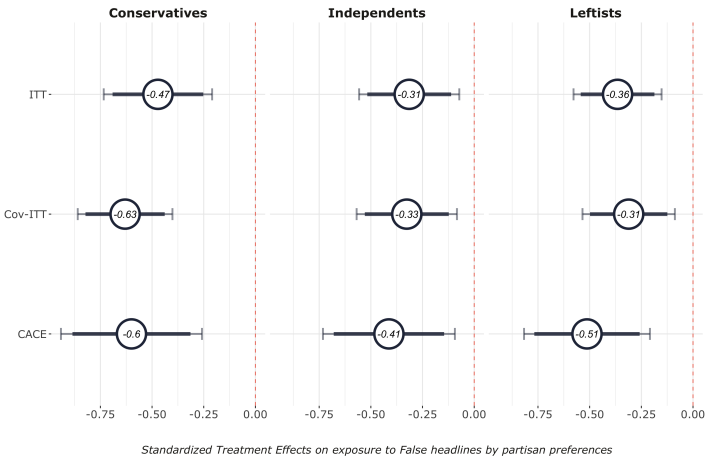


Figure 7 Treatment effects of WhatsApp Multimedia Deactivation on exposure to pro-attitudinal and counter-attitudinal false items conditional on users' voting preferences. The figure presents standardized coefficients with their corresponding 95% and 90% confidence intervals based on robust standard errors.

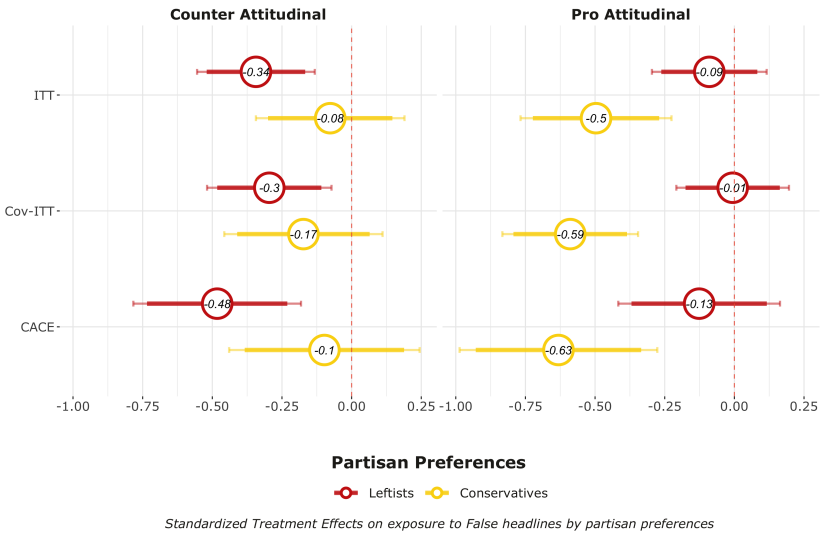


Figure 8 Treatment effects of WhatsApp Multimedia Deactivation on Truth Discernment. The figure presents standardized coefficients with its corresponding 95% and 90% confidence intervals based on robust standard errors.

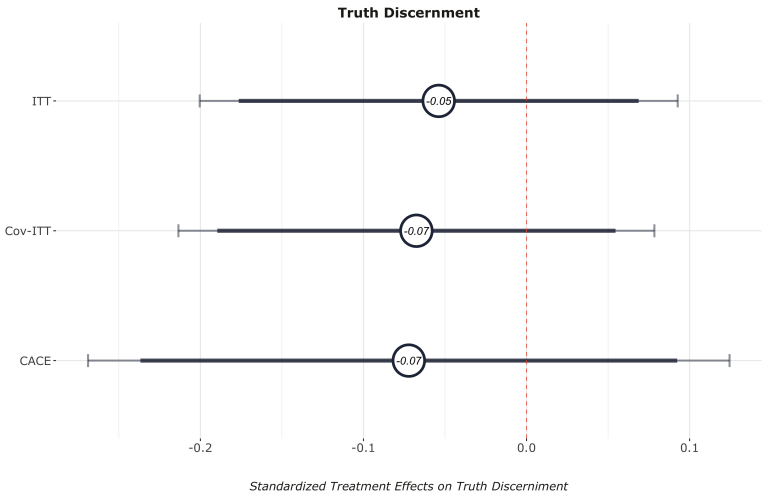


Figure 9 Treatment effects of WhatsApp Multimedia Deactivation on *False Rumors Accuracy* and *True News Accuracy* conditional on users' voting preferences. The figure presents standardized coefficients with their corresponding 95% and 90% confidence intervals based on robust standard errors.

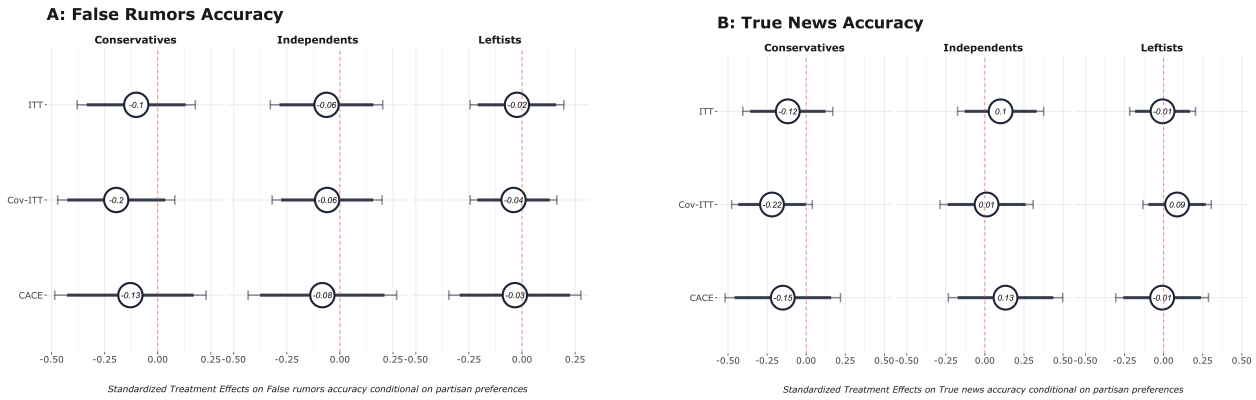


Table 7 Regression Models: Deactivation Treatment Effects on Self-Reported Exposure to Misinformation on WhatsApp. The outcome variable is a nominal scale measuring self-reported exposure to misinformation. Results indicate robust negative decrease in self-reported exposure to false rumors, as depicted in the headline task.

	ITT	CACE
Intercept	3.450***	3.450***
	(0.103)	(0.103)
Treatment	-0.452**	-0.606**
	(0.145)	(0.194)
Num.Obs.	732	732
R2	0.013	0.016
R2 Adj.	0.012	0.014
RMSE	1.96	

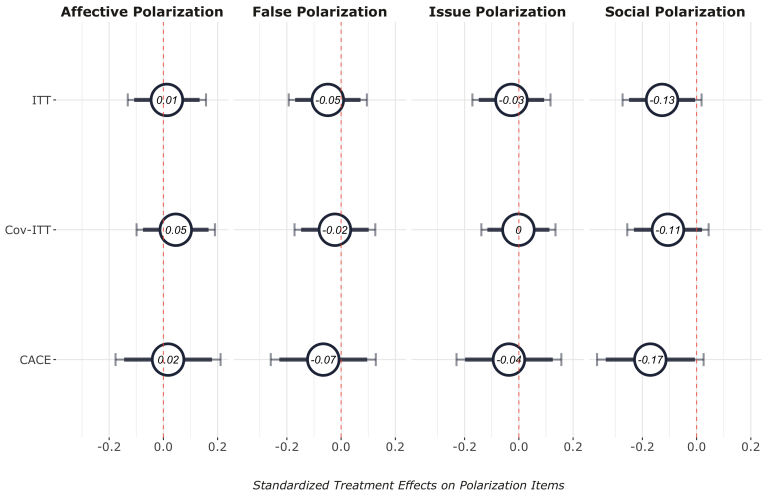
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note:

Robust standard errors in Parentheses

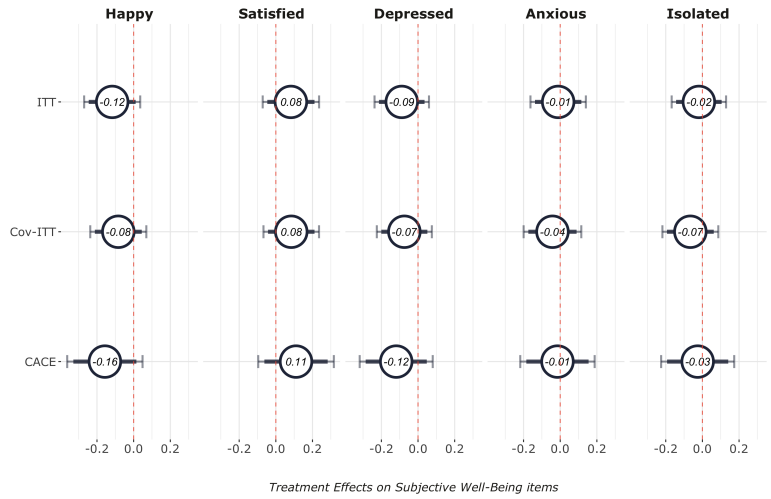
6 Additional Results: Polarization

Figure 10 Treatment effects of WhatsApp Multimedia Deactivation on four measures of polarization.



7 Additional Results: Subjective Well-Being

Figure 11 Treatment effects of WhatsApp Multimedia Deactivation on five subjective well-being items. 1



8 Heterogeneous Effects (Age, Digital Literacy, Using WhatsApp for Politics), Effects on Trust and Substitution Effects

We pre-registered three primary moderators on which we commit to reporting heterogeneous dynamics of the deactivation effects. Those are: participants' age, levels of digital literacy, and use of WhatsApp to share and receive information about politics. For age, we use the pre-treatment survey question. For digital literacy, we create an index summing the z-score across four items measuring digital literacy from the post-treatment survey[1] To measure participants' usage of WhatsApp for politics, we asked participants how often they use WhatsApp to send/receive information about politics and elections to their friends, using a nominal scale with the following values: At least ten times a day, Several times a day, About once a day, three to six days a week, one to two days a week, Every few weeks, Don't Know. For the models presented below, we converted this scale to numeric and considered "Don't Know" as a missing value. This section also presents additional analysis of the deactivation on trust in the media and self-reported substitution effects from the deactivation.

[1]The four items are the following: *a) I prefer to ask friends how to use any new technological gadget instead of trying to figure it out myself; b)I feel like the internet is a part of my daily life; c) Using information technology makes it easier to do my work; d) I often have trouble finding things that I've saved on my computer.* We asked participants for their agreement with these statements using a 1-7 scale.

Table 8 Regression Models: Heterogenous ITT effects of the Deactivation conditional on Digital Literacy Score

	False Rumors Exposure	True News Exposure	False Rumors Accuracy	True News Accuracy	Polarization Index	Subjective Well-Being Index
Treatment	-0.424***	-0.282***	-0.069	-0.009	-0.029	0.154
	(0.077)	(0.065)	(0.082)	(0.080)	(0.159)	(0.273)
Digital Literacy	0.017	-0.001	0.074**	0.013	0.152**	-0.009
	(0.025)	(0.022)	(0.023)	(0.025)	(0.050)	(0.081)
Treatment x Digital Literacy	-0.048	-0.023	-0.008	0.014	0.036	-0.180
	(0.035)	(0.029)	(0.034)	(0.036)	(0.071)	(0.122)
Num.Obs.	660	660	660	660	660	660
R2	0.163	0.140	0.188	0.104	0.161	0.126
R2 Adj.	0.122	0.097	0.148	0.060	0.120	0.083
RMSE	0.94	0.79	1.01	0.98	1.97	3.34

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

Note: Robust standard errors in Parentheses. All models use the Covariate-Adjusted ITT estimator.

Table 9 Regression Models: Heterogenous ITT effects of the Deactivation conditional on Age

	False Rumors Exposure	True News Exposure	False Rumors Accuracy	True News Accuracy	Polarization Index	Subjective Well-Being Index
Treatment	-0.204	-0.374*	-0.100	-0.112	-0.077	-0.177
	(0.199)	(0.160)	(0.206)	(0.200)	(0.402)	(0.646)
Age	-0.051	-0.050	0.017	0.008	0.102	0.357+
	(0.060)	(0.046)	(0.057)	(0.059)	(0.116)	(0.196)
Treatment x Age	-0.097	0.042	0.003	0.043	-0.003	0.156
	(0.077)	(0.066)	(0.084)	(0.079)	(0.163)	(0.272)
Num.Obs.	660	660	660	660	660	660
R2	0.162	0.138	0.169	0.102	0.132	0.120
R2 Adj.	0.122	0.097	0.130	0.060	0.090	0.078
RMSE	0.94	0.79	1.02	0.98	2.00	3.35

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

Note: Robust standard errors in Parentheses. All models use the Covariate-Adjusted ITT estimator.

Table 10 Regression Models: Heterogenous ITT effects of the Deactivation Conditional on WhatsApp Usage for Politics

	False Rumors Exposure	True News Exposure	False Rumors Accuracy	True News Accuracy	Polarization Index	Subjective Well-Being Index
Treatment	-0.313*	-0.305**	-0.298*	-0.119	-0.014	0.172
	(0.137)	(0.103)	(0.144)	(0.134)	(0.262)	(0.474)
WhatsApp usage for Politics	0.036	0.060+	-0.090*	0.024	-0.065	-0.215+
	(0.042)	(0.034)	(0.041)	(0.040)	(0.077)	(0.122)
Treatment x WhatsApp usage for Politics	-0.075	-0.015	0.096*	0.023	0.024	0.084
	(0.051)	(0.039)	(0.048)	(0.048)	(0.097)	(0.161)
Num.Obs.	517	517	517	517	517	517
R2	0.137	0.137	0.187	0.110	0.135	0.133
R2 Adj.	0.083	0.085	0.137	0.057	0.084	0.081
RMSE	0.99	0.83	1.00	0.98	1.98	3.27

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

Note: Robust standard errors in Parentheses. All models use the Covariate-Adjusted ITT estimator.

Table 11 Regression Models: Covariate Adjusted Models for Deactivation Treatment Effects on Trust Measures. The survey data uses a seven-point nominal scale measuring trust in various institutions and news media. The outcome for the models uses a standardized z-score for the indicator trust measures, and the sum of the z-scores for the index. Robust standard errors are shown in parentheses. Our results do not recover any statistically significant effect of the deactivation on standard measures for trust

	Trust Score	Trust Government	Trust Congress	Trust Electoral Authorities	Trust Globo (TV News Media)	Trust News Outlets
Intercept	-1.146	-0.695+	0.103	-0.359	-0.027	-0.168
	(1.226)	(0.401)	(0.417)	(0.355)	(0.348)	(0.342)
Treatment	-0.025	-0.027	-0.002	-0.040	0.014	0.031
	(0.211)	(0.065)	(0.068)	(0.060)	(0.054)	(0.058)
Num.Obs.	660	660	660	660	660	660
R2	0.440	0.320	0.259	0.423	0.508	0.433
RMSE	2.65	0.83	0.86	0.76	0.70	0.75

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

Table 12 Regression Models: Covariate Adjusted Models for Substitution effects of the Deactivation Treatment. The survey data uses a five-point nominal scale asking how much less/much more time participants spent in a set of online and offline activities during the weeks of the experiment. Our results do not uncover a substitution of WhatsApp for other online activities. We find participants rely more on TV News as a consequence of the deactivation.

	Other Social Media	Reading News Online	Watching TV	Offline with Friends	Away from my Phone
Intercept	3.271***	1.685**	5.636***	4.233***	3.850***
	(0.548)	(0.522)	(0.519)	(0.495)	(0.509)
Treatment	0.124	-0.063	0.215**	0.012	0.139
	(0.086)	(0.084)	(0.079)	(0.079)	(0.085)
Num.Obs.	660	660	660	660	660
R2	0.065	0.058	0.107	0.066	0.050
RMSE	1.09	1.06	1.01	0.99	1.06

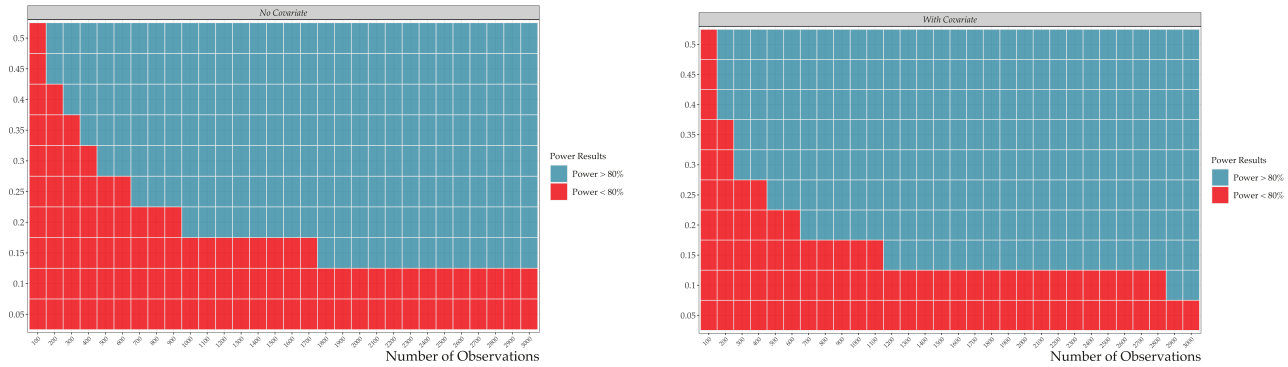
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: Robust standard errors in Parentheses

9 Power Analysis

We estimated and pre-registered the sample size for our design, taking as a baseline the effect size of previous deactivation studies (Asimovic et al., 2021; Allcott et al., 2020). We used the *DeclareDesign* Framework (Blair et al., 2019) to simulate the sample size needed. As in Asimovic et al. (2021); Allcott et al. (2020), we use standardized effects for the dependent variables and find an average effect of 0.2 standard deviations as the average effect size in the deactivation literature for news knowledge and polarization indexes (our primary hypotheses). We simulate the power statistics using two different estimators, ITT and COV-ITT.

Figure 12 Power Analysis



Notes: Power analysis using simulations with DeclareDesign. The upper figure presents models without covariate adjustment. The bottom figure presents results with adjustments. The model assumes a 0.2 correlation coefficient between the outcome and covariates.

With no controls, a sample size of 1000 participants would be well-powered (power above 0.8) to identify small effects, as small as 0.2 standard deviations, from the baseline from previous work. Using a COV-ITT with at least a 0.2 correlation between dependent variables and controls, a sample size of 700 participants would allow us to detect a 0.2 standardized effect. Therefore, our experiment is well-powered to detect effects of a similar size to the previous deactivation studies when using the COV-ITT estimator. Figures 12 present the results of the power analysis.

10 Effects of Exposure on Beliefs for Misinformation

In this section, we present models regressing False Rumors Accuracy on False Rumors Exposure for all items. These models show how exposure shapes belief accuracy when measured as a self-reported endogenous variable. However, when using our well-identified reduction in exposure to misinformation, these effects on accuracy disappear.

Table 13 Regression Models: Models regressing False Rumors Accuracy on False Rumors Exposure for all items. The first column presents a simple OLS model with no controls. The second column presents the Covariated-Adjusted Model. The third column presents an instrumental variable estimation with exposure as the endogenous variable. The results indicate exposure strongly predicts belief in misinformation. However, the effects disappear when endogenous components are removed using the instrumental variable model.

	ITT Belief Misinfo ~ Exposure	Cov-ITT Belief Misinfo ~ Exposure	Belief Misinfo ~ Exposure | Deactivation
Intercept	0.180***	0.048	0.402*
	(0.009)	(0.105)	(0.185)
Exposure to False Rumors	0.229***	0.223***	-0.190
	(0.024)	(0.025)	(0.171)
Num.Obs.	2928	2640	2640
R2	0.049	0.072	-0.080
R2 Adj.			-0.092
RMSE	0.41	0.40	
Std.Errors	by: q_email	by: q_email	by: q_email
Controls	no	yes	yes

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

Robust standard errors in parentheses

11 Multiple Hypothesis Testing

Table 14 Unadjusted and Adjusted P-Values Testing Each Hypothesis

	H1	H2	H3	H4	H5
Unadjusted P-Value	0	0.7509	0.7509	0.2343	0.7555
P-Value (FDR Adjusted)	0	0.7555	0.7555	0.58575	0.7555

Table 15 Unadjusted and Adjusted P-Values for Polarization Outcomes

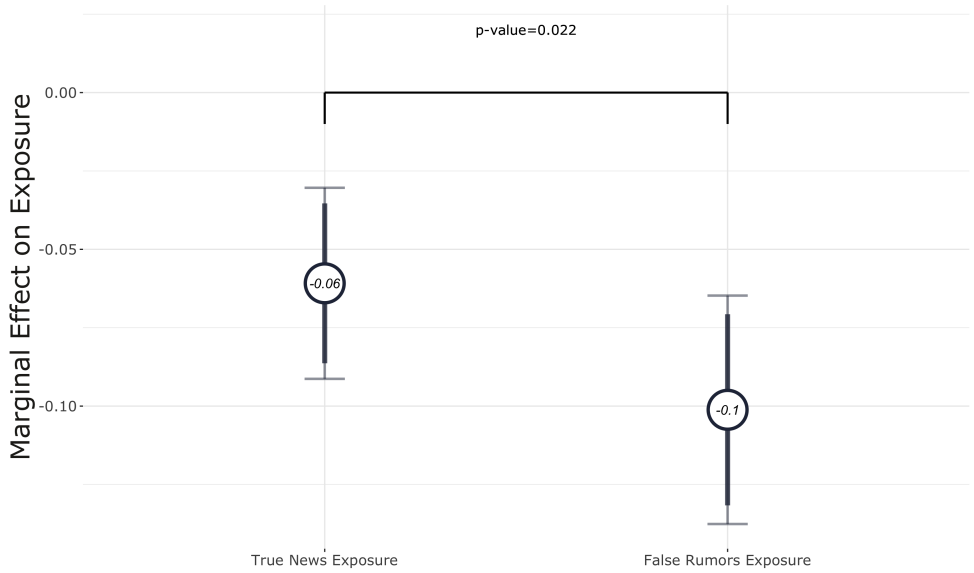
	Polarization Index	False Polarization	Affective Polarization	Social Polarization	Policy Polarization
Unadjusted P-Value	0.23	0.50	0.86	0.09	0.71
P-Value (FDR Adjusted)	0.59	0.84	0.86	0.44	0.86

Table 16 Unadjusted and Adjusted P-Values for Subjective Wellbeing Outcomes

	Index	Happy	Depressed	Anxious	Isolated	Satisfied
Unadjusted P-Value	0.76	0.13	0.24	0.88	0.80	0.30
P-Value (FDR Adjusted)	0.88	0.59	0.59	0.88	0.88	0.59

12 Item-Level Analysis

Figure 13 Item-Level Marginal Effects of Treatment on Exposure conditional on True/False items. The marginal effects come from item-level models regressed on the participant's treatment condition interacted with the item's veracity (true news or false rumors). Standard errors are clustered at the participants' level.



13 Descriptive of WhatsApp Usage in the Final Sample

In this section, we provide descriptive statistics of WhatsApp usage among the participants in our experiments. Figure 14 shows that almost half of the sample uses WhatsApp for over 4 hours per day; and over 80% use it for at least one hour per day. Figure 15 shows that almost 60% in the pre-treatment survey reported receiving media related to politics and elections on WhatsApp at least once every day, and Figure 16 shows that about 95% use WhatsApp at least once a day to communicate with family and friends, and between 50% and 70% use it at least once a day to consume political content. These numbers indicate that our sample represents well the high penetration and high-levels of activity of WhatsApp users in Brazil *(27)* and particularly the high prevalence of political content in the context of the election.

Figure 14 Pre-treatment distribution of self-reported daily WhatsApp usage.

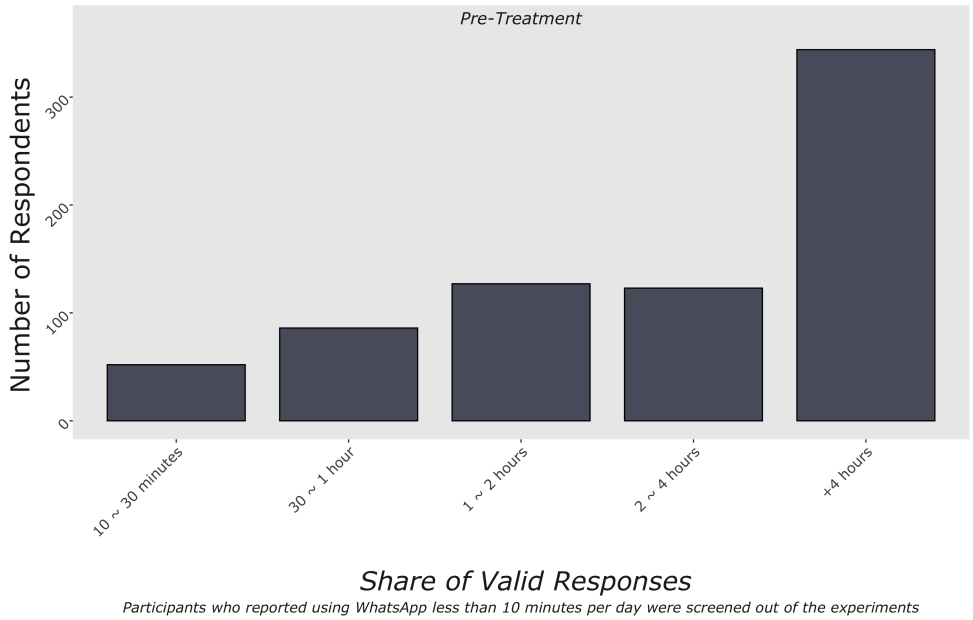


Figure 15 Pre-treatment distribution of self-reported consumption of media on WhatsApp related to politics and elections

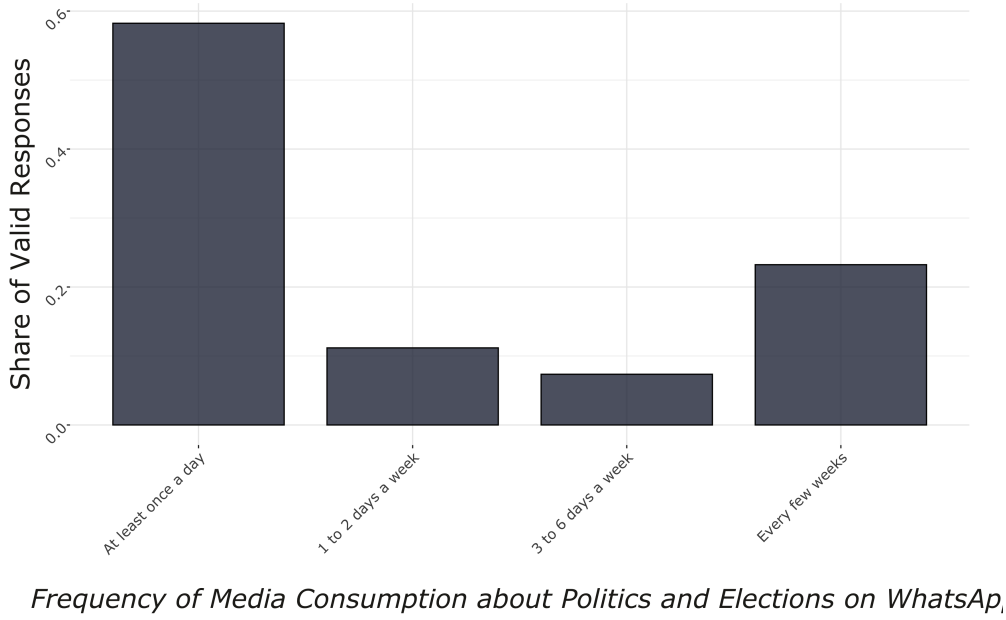
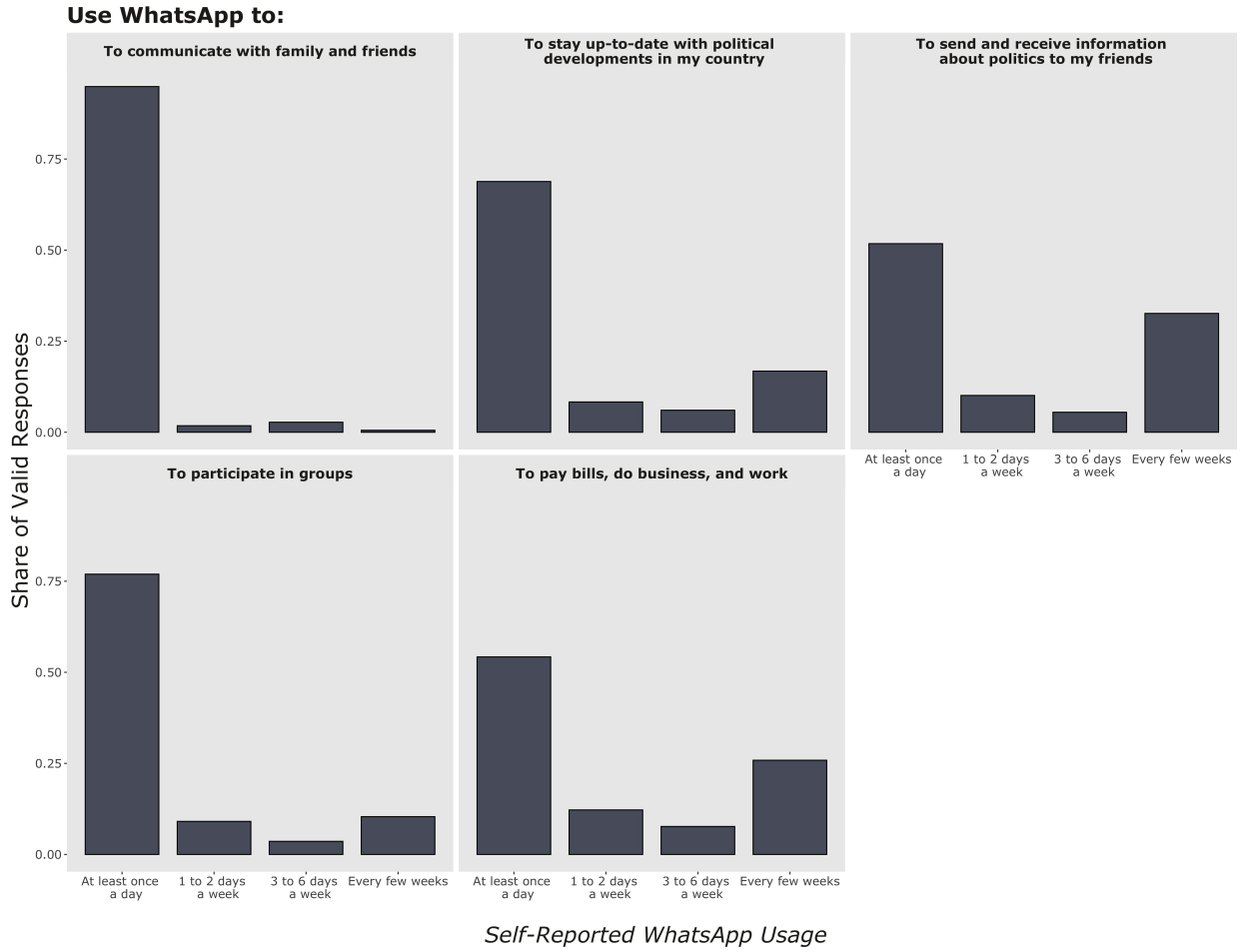


Figure 16 Pre-treatment distributions of self-reported WhatsApp usage for a variety of personal, professional, and political activities.



14 Deviations from the Pre-Analysis Plan

In this section, we report the deviation from the Pre-Analysis Plan [REDACTED]. First, we switch the order of the hypotheses. In the Pre-analysis plan, our first hypothesis related to belief accuracy, and the second hypothesis focused on reducing exposure to misinformation. In the manuscript, we swap the order of the hypothesis since we believe our results are easier to read, first focusing on changes in exposure and then on theoretical expectations driven by this informational shock. In addition, we slightly edited the text of the hypotheses to improve readability.

References

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

Aruguete, Natalia, Ernesto Calvo and Tiago Ventura. 2021. "News sharing, gatekeeping, and polarization: A study of the# Bolsonaro Election." *Digital journalism* 9(1):1–23.

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).

Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2019. "Declaring and diagnosing research designs." *American Political Science Review* 113(3):838–859.

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

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