The Effect of Social Media on Voters: Experimental Evidence from an Indian Election

Kevin Carney*

September 15, 2023

Abstract

This paper uses a field experiment to study the effect of social media on voters in a large election in Tamil Nadu, India. I randomly invited study participants to join chat groups organized by political parties on WhatsApp, India's most-used social media platform. Unlike Facebook and Twitter, WhatsApp groups are relatively small, and users' feeds are neither curated by algorithms nor moderated by the platform. I find that joining a group increases political knowledge, improving participants' ability to distinguish true from false news. Moreover, the groups have an effect on political preferences, pushing participants toward the party affiliated with the WhatsApp group they joined. This effect is driven by moderates who were indifferent between parties at the baseline. To disentangle the effect of direct party messaging from the effect of messaging between individual users, I designed a second treatment arm that exposed participants to the posts of party officials but not the posts and replies of other group members. This reveals that the horizontal exchange between individuals is key: the treatment effects of party messaging alone are smaller than those of the full groups. I provide evidence on possible mechanisms underlying this difference. Participants assigned to the full groups both receive more messages and pay more attention to the messages they receive.

*The University of Chicago: kec@uchicago.edu. I thank to Emily Breza, Michael Kremer, David Yang, Gautam Nair, and especially Gautam Rao for indispensable advice and guidance. Vasanthi Pillai, Niveditha Lakshmi, Rakesh Pandey, and Alosias A provided exceptional research assistance and project management. This project was possible thanks to funding from the Harvard Kennedy School's doctoral program. This project was approved by the Institutional Review Boards at Harvard University and the Institute for Financial Management and Research (IFMR). The experiment is listed in the AEA RCT Registry (AEARCTR-0007371).

1 Introduction

Mobile phone ownership has exploded in emerging economies. This technology has eased information frictions in contexts ranging from price dispersion across markets (Jensen, 2007) to public health knowledge during the pandemic (Banerjee et al., 2020). Recently, social media has become a forum for political conversations and a tool for political parties and politicians to reach voters. Such uses of social media may expand the information available to voters in developing democracies, where information constraints can limit citizens' ability to enforce good governance (Pande, 2011). Yet information shared on social media can be incorrect or selected. Social media favors short-form communication and may be uniquely suited to spread oversimplified or misleading narratives (Guriev et al., 2021). The effects of social media on voter knowledge, preferences, and behavior are therefore unclear.

This paper addresses two main research questions. First, how does political parties' use of social media affect voters in a developing country setting? Second, does the effect of social media depend on interpersonal engagement between users? A defining feature of social media is that users engage in horizontal communication with each other. Traditional media involves vertical communication, where information comes from a common source with limited interaction. By contrast, social media involves elements of both vertical and horizontal communication (Zhuravskaya et al., 2020). Across platforms, political parties and politicians use social media to broadcast mass messages. But individual users can also post and reply. Is social media just a tool for parties and politicians to reach more voters? Or does the interpersonal engagement between users drive social media's effects?

To answer these questions, I conducted a field experiment in India on WhatsApp. WhatsApp is a messaging application and India's largest social media network. With over half a billion Indian users, it is a common medium for political conversations. Such conversations often occur in groups of up to 256 people (WhatsApp's cap). Political parties have taken advantage of this technology to reach voters. In advance of elections, parties cre-

ate thousands of localized WhatsApp groups wherein members can share political content. Group administrators also push curated campaign posts to the groups. Recent estimates suggest that one in six WhatsApp users in India have participated in a political group on the platform (Kumar and Kumar, 2018). Political parties use WhatsApp groups to campaign in many of the world's largest democracies, including Indonesia, Brazil, Mexico, and the Philippines (Garimella and Eckles, 2020).

I studied WhatsApp groups created in the context of statewide elections in Tamil Nadu, India. These were large elections for the state Legislative Assembly, with 46 million citizens casting their vote in April 2021. Both of the two leading parties created political WhatsApp groups organized at the constituency level.[1] I identified 25 constituencies around Chennai where both parties posted public invitation links for interested people to join their groups. Groups contained 94 members on average before my intervention.

In the experiment, I randomly offered participants links to join one of these WhatsApp groups. To isolate the effect of interpersonal engagement between users, I designed a second treatment arm. This treatment exposed participants to the vertical messaging from the parties but not the horizontal messaging from other group members. For each original party-organized WhatsApp group, I created a new corresponding group. However, the group settings prevented members themselves from posting. Instead, I created an automation program that identified the posts from party administrators in the original groups and then forwarded these messages to the new groups. In this way, the new groups exposed participants to the exact party messaging they would have seen in the original groups, but not the posts and replies of other members.

I randomized participants to the full WhatsApp groups, party-content-only groups, or a control group. I recruited a sample of 3,056 WhatsApp users using Facebook ads targeted at the study constituencies during the month before the election. The ads redirected participants

[1]Seats in the legislature are assigned at the constituency level. Tamil Nadu has 234 constituencies. Constituencies represent more than 100,000 voters, and each party created many WhatsApp groups per constituency. Thus, it is unlikely that group members knew each other outside of the group.

to a brief online baseline survey. At the end of the survey, participants saw a link to a WhatsApp group for one of the two leading parties. The party was randomly assigned, and the group corresponded to the participant's stated constituency. Of participants who reached the invitation link, 49% clicked to join the group offered. Upon clicking the link, participants were randomly assigned to a treatment condition: full group, party-content-only group, or control. Those assigned to a full or party-content-only group were automatically redirected to that group in WhatsApp. Those assigned to the control were instead redirected to the final page of the survey. This randomization allows me to estimate unbiased treatment effects for those individuals who decided to click the link for the randomly assigned party. Approximately four weeks later, shortly after the election, I conducted an endline survey. Of experimental participants, 83% completed the endline, and attrition was balanced across the three treatment arms.

My first result is that political WhatsApp groups increase knowledge about political news. Participants in full groups were better able to distinguish true from false news. This result is striking considering concerns about political misinformation on WhatsApp in India and on social media more generally. I measured knowledge using an endline quiz. Participants viewed a list of nine news headlines and responded with how confident they were that the stories were true or false. Each participant saw three types of headlines: three that were true, taken from the news; three that were rumors circulating online, debunked by prominent fact-checkers; and three that were false but not circulating online, written by research assistants in Chennai. The nine headlines were drawn at random from a longer list and appeared in random order. Scores on the quiz increased by 0.17 standard deviations (SE 0.07) in the full groups relative to the control. This increase came mainly from treated participants' confidence in true headlines about the assigned party, which increased by 0.48 standard deviations (SE 0.18) relative to the control. Belief in rumors and false headlines decreased, but point estimates are small and not significantly different from zero.

My second result is that the full groups have a significant average effect on political preferences, pushing participants toward the assigned party. This effect comes mainly from participants who identified as moderate at the baseline. I measured preferences by indexing participants' ratings of the two main parties and their leaders. The relative rating of the assigned party increased by 0.20 standard deviations (SE 0.08) for participants in the middle quintile of the baseline preference distribution. Treatment effects at other points of the distribution have small point estimates that are not significant at conventional levels. Importantly, I can rule out any meaningful backlash effects among participants assigned to the group of a party they did not like at the baseline. I also do not find any effect on affective polarization—negative feelings toward members or supporters of the opposing party—using standard survey measures of partisan animus (Iyengar and Krupenkin, 2018; Levy, 2021).

My third result is that horizontal communication between group members is key to the groups' treatment effects. Across all main outcomes, the treatment effects of party messaging alone are consistently smaller and less significant than those of the full groups. Party messaging alone has no significant effect on knowledge or political preferences. Using an index of these main outcomes, I show that the difference between the aggregate effects of the full groups and party messaging alone is statistically significant. This difference suggests that the defining feature of social media—horizontal communication between peers—is precisely what makes social media groups influential in this context.

I find that exposure to horizontal communication increases engagement in the WhatsApp groups. Detailed administrative data from WhatsApp highlight two mechanisms that distinguish the full groups from the party-content-only groups: message volume and attention. These data allow me to observe all posts that participants were exposed to and how quickly posts were viewed (using WhatsApp's "seen by" feature). The full groups had substantially higher volume, receiving on average 113 posts per day, of which only seven came from party officials. However, using (nonrandom) heterogeneity between groups, I find no significant differences in treatment effects by message volume (party or total) in the full

groups. Ex ante, the relationship between message volume and attention is unclear: more messages could increase engagement but could also induce fatigue, leading participants to mute or leave a group. The administrative WhatsApp data show that the former effect dominates. Despite the higher volume of messages, full-group participants were more likely to view a message they received than party-content-only participants. Consistent with this, participants in the full groups reported spending 16 minutes more on WhatsApp each day than participants in the party-content-only condition or the control. These results demonstrate that the horizontal communication in WhatsApp groups increases user engagement.

There are other possible explanations for the divergent treatment effects of the full group and party-content-only conditions. People may draw different inferences from a message depending on its sender—party or individual—or they may be drawn in by the opportunity to post to the full groups themselves. One limitation to this study is that it does not provide exogenous variation in all the possible mechanisms that differentiate the two main treatment conditions, nor would it be practical to do so. These potential mechanisms are not mutually exclusive, and there could be interactions between them. Understanding the foundations of how people's preferences and beliefs update in response to social media interactions is an important avenue for future research.

Taken together, the results of this paper constitute encouraging evidence on the role of social media in developing democracies. Political groups on WhatsApp improved voter knowledge, making participants better able to distinguish true from false news. While the groups pushed members' preferences toward the assigned party, heterogeneity analysis suggests that the groups' content was more tailored to moderates than extremists, and affective polarization did not increase. Considering widespread concern about the effects of social media on the democratic process, these results paint a more optimistic picture. Future research on other types of social media groups and political environments will help assess the generalizability of these findings.

This paper contributes to the development economics literature on voter information. This literature largely focuses on information shared from a common source through mass media. By contrast, information shared between peers on social media may be of lower quality, making learning more difficult. Existing research finds that information disseminated by mass media can empower voters to hold politicians accountable and make decisions that more closely align with their preferences (Ferraz and Finan, 2008; Da Silveira and De Mello, 2011; Banerjee et al., 2020). Experiments have identified interventions to increase voter knowledge, such as SMS messages (Aker et al., 2017; George et al., 2020; Marx et al., 2021), politician report cards (Banerjee et al., 2011), and film (Ravallion et al., 2015). One result from this literature is that information campaigns with an interactive component may be particularly effective, as in the town hall meetings studied by Fujiwara and Wantchekon (2013) or the debates studied by Bidwell et al. (2020). This study is unique because its design separates the effects of one-way messaging from party to voter and two-way messaging between voters. I show that social media can be a valuable source of political information, and it may be more engaging than vertical communication alone.

This paper also contributes to growing experimental literature on the political effects of social media. This literature focuses mainly on the United States and platforms such as Facebook and Twitter (Levy, 2021; Bursztyn et al., 2021; Allcott et al., 2020; Jiménez-Durán, 2021). My study contributes to the literature by providing new evidence on a large class of social media platforms: messaging applications. Messaging applications differ from other social media platforms in that no algorithm dictates which posts receive the most attention. All posts appear in reverse-chronological order, and the groups are relatively nonhierarchical. Social media algorithms that instead account for message engagement may disproportionately feature surprising, provocative, or inflammatory content. This is less of a concern on messaging applications like WhatsApp, where algorithms do not govern message order. Conversely, WhatsApp's content is effectively unregulated and unmoderated by the platform due

to its end-to-end encryption. WhatsApp does not fact-check or censor messages in any way, possibly making misinformation even easier to spread.

Most relevant from this literature is Allcott et al. (2020), where the authors pay Facebook users to deactivate their accounts in the month before the midterm elections in the U.S. They find that leaving Facebook decreases news knowledge and political polarization. This is directionally consistent with my results on the impact of WhatsApp on knowledge and political preferences. Other experiments study the impact of specific features of Facebook, such as news pages users can follow (Levy, 2021) or an "I Voted" button (Bond et al., 2012), finding impacts on affective polarization and voting behavior, respectively. My research design builds on this research by isolating the feature of social media that differentiates it from traditional media: the ability of individual users to communicate with each other. This paper also contributes to this literature by measuring the effect of social media in under-studied contexts. There is limited research on the impact of social media on voters in developing democracies, where social media use is increasing rapidly. This paper aims to contribute to our understanding of how social media affects voters across varied contexts.

2 Context

The use of WhatsApp in political campaigns is not unique to Tamil Nadu, where this study took place. In this section, I describe WhatsApp's role in elections around the world and characterize the WhatsApp groups created for Tamil Nadu's 2021 elections.

2.1 WhatsApp Elections

WhatsApp is incredibly popular in emerging economies. WhatsApp is India's most-used social media platform, with an estimated 530 million users (Bharadwaj, 2021). Many credit this popularity to WhatsApp's accessibility. Relative to SMS, WhatsApp is inexpensive. The application is free and requires minimal data. It can be used on all smartphones, the costs of which have fallen in the past decade. Providers in locations where WhatsApp is popular offer budget smartphones and data plans with internet access exclusive to WhatsApp

use. WhatsApp is also accessible to people with limited literacy and digital literacy. The application is easy to use and does not strictly require reading or typing skills. People who cannot read or type send and receive voice recordings, photos, and videos on WhatsApp. Consequently, WhatsApp users are drawn from across the education and age distribution.

This makes WhatsApp a powerful tool for political campaigns. In India, the national ruling party, the Bharatiya Janata Party (BJP), pioneered the use of WhatsApp in campaigning. They developed a vast network of WhatsApp groups, with tens of thousands of "IT cell" volunteers tasked with pushing content to the groups (Murgia et al., 2019). This was central to the party's strategy, with the BJP's social media head declaring the 2019 general election a "WhatsApp election," a term echoed in the national and international press (Perrigo, 2019). Based on the BJP's success on social media, other parties have adopted a similar strategy. The use of WhatsApp in political campaigns has spread around the world, particularly in developing democracies, giving parties access to voters who are otherwise difficult to reach (Renno, 2019).

2.2 Tamil Nadu 2021 Election

I conducted the study in the month before Tamil Nadu's 2021 Legislative Assembly election. Tamil Nadu is India's sixth-largest state, with 63 million registered voters, and the Legislative Assembly is the state's governing body. At stake in this election were 234 constituency seats in the assembly and the state's Chief Minister. The election was a contest between Tamil Nadu's two leading regional parties: the incumbent All India Anna Dravida Munnetra Kazhagam (AIADMK) and the challenging Dravida Munnetra Kazhagam (DMK). These parties have traded control of the state legislature since 1967. Most other parties contesting seats formed alliances with either the AIADMK or the DMK. This study focuses on these two leading parties. The AIADMK is characterized as right-leaning, allying with the Bharatiya Janata Party (BJP) in national politics, and the DMK leans more to the left, allying with the Indian National Congress. The 2021 election was consequential because it was the first election after both parties' longtime leaders died in 2016 and 2018, so the election represented a new era

of Tamil politics. The voter turnout rate of 73% indicates the perceived importance of this election.

2.3 The WhatsApp Groups

The AIADMK and DMK created thousands of WhatsApp groups leading up to the 2021 elections. Both parties organized their groups at the constituency level, with multiple groups per constituency. The parties recruited people to these groups using hyperlinks that WhatsApp users could click to view and join a group. These links were also shared on social media and the parties' websites. Several months before the election, research assistants in Chennai and I searched social media and identified 25 constituencies in the Chennai area where links for groups organized by both major parties were publicly available. These groups existed independent of the study, created by party organizers. The groups in the study represent a small subsample of all groups created by political parties to campaign in this election.[2]

WhatsApp groups work similarly to other group messaging services, where all members of a group receive all posts to the group as they come in. Unlike social media platforms that rely on algorithms to feature content, WhatsApp group messages appear in the order they are posted, with new messages added to the bottom of the group thread. There are no "like" buttons as on Facebook and Twitter.[3] WhatsApp puts a cap on group size at 256 members. The groups in the study were large, with 94 members on average prior to the intervention. Of these members, a handful in each group were party volunteers designated as the group's administrators, and the rest were members of the general public who opted in through an invitation link. Group members were not likely to know each other outside of the groups.

[2]Technological and team constraints prevented us from identifying and including all groups created by the parties. The parties often relied on sharing group links through WhatsApp networks rather than more public social media platforms like Facebook and Twitter. Additionally, the AIADMK sometimes shared links by constituency using location services, limiting our team's ability to collect invitation links outside the Chennai area. However, the groups I observed that were not included in the study are not noticeably different from those included in the study. The groups used in the study are also similar to political groups organized by other parties in other states of India.

[3]WhatsApp users can reply to messages and sometimes reply with emojis to react to a post. Unlike Facebook and Twitter, these reactions are not aggregated, appearing as individual posts.

The groups' content settings allowed all members to post freely. The party administrators did not moderate this (horizontal) communication between group members.[4] Group members posted various content, including images (memes, political cartoons, infographics), written messages, and external links to articles or videos. Party administrators engaged in the groups using (vertical) communication from the central party. Party administrators often coordinated this communication across groups, posting a common message to all groups simultaneously. The administrators did not typically engage beyond these posts, even if group members replied. Party posts were generally professionally made images and videos with accompanying text. This content was similar to campaign advertising on traditional media.

The combination of interpersonal messaging between group members and direct messaging from the parties is a fundamental feature of the groups. In the economics literature on social media, this mix of horizontal and vertical communication distinguishes social media from traditional media (Zhuravskaya et al., 2020). The parties used the groups as a form of mass communication, pushing content to group members vertically, similar to their newspaper, radio, and television advertising, with little to no interaction. Group members could post content themselves and see the posts and replies of others, communicating horizontally. Other platforms such as Facebook and Twitter similarly combine horizontal and vertical communication as means for political engagement. Does social media primarily function as a new mass media tool, or do the interactions between individuals drive social media's effects? I describe how I parsed the effects of these features in the following section.

3 Research Design

Using the WhatsApp group links, I conducted a randomized experiment to study the effects of the groups on their members. This section describes the experimental design.

[4]Individual users sometimes self-moderated the groups, discouraging other members from posting irrelevant or counter-attitudinal messages. In practice, these types of messages were rare, limited to occasional spam advertisements.

3.1 Treatment Arms

To measure the effect of the groups on their members, I randomized participants to join a political WhatsApp group or to a control group. Section 3.2 describes the randomization process in detail.

To separate the effect of vertical communication by the political parties from horizontal communication by group members, I created a second treatment arm. This arm exposed participants to the party messaging they would have observed in the WhatsApp group for their constituency (vertical communication), but not the posts of other general group members (horizontal communication). To do this, I created a new WhatsApp group corresponding to each party-organized group. In the new group, the content settings prevented participants themselves from posting. Instead, they could simply observe the group administrator's posts, which were the posts of party officials in the corresponding "full" group. I developed an automated web driver program in Python that identified the posts coming from party organizers and forwarded those messages to the new group as they came in. Crucially, the automation did not forward any other messages. Participants in this "party-content-only" condition were exposed to the exact vertical communication by the parties that they would have observed in the full group without seeing the horizontal communication between other group members.

This yields three distinct treatment conditions: (1) a "full-group" condition, (2) a "party-content-only" condition, and (3) a control.

3.2 Randomization

Participants were recruited into the study via Facebook, as described in Section 3.4. They first completed a brief, self-administered survey on Qualtrics to measure baseline political preferences and demographic variables. Next, they were provided with a link to join a WhatsApp group for their constituency for a randomly assigned political party (AIADMK or

DMK).[5] Participants did not receive an incentive to join the group. Instead, as in daily life, they could freely choose to sign up or not.[6]

The obvious challenge posed by the above design—and shared by many field, especially online field experiments—is imperfect treatment compliance. While participants are provided links to join randomly assigned groups, they select whether to comply with the invitation. Effects of the treatment on the treated (TOT) can be recovered by instrumenting for group membership with treatment assignment. However, low compliance can dramatically decrease statistical power in such a design: the required sample size for an experiment increases inversely to the square of the compliance rate. Moreover, differential attrition between those who choose to join the assigned group and those who decline poses a serious threat to such a design.

Therefore, I designed the experiment to randomize treatment assignment *among compliers*. To do this, I exploited the survey technology, which allowed me to randomize conditional on clicking the group invitation link that was displayed. In effect, this screened out "never-takers" prior to randomization. First, participants were shown an invitation link to a randomly assigned party's group (corresponding to their constituency). Then, participants who clicked the link were randomly redirected either to join the full group in WhatsApp, to join the party-content-only group in WhatsApp, or to the control condition without any group, in which case they were simply redirected to the last page of the survey. This second randomization, conditional on clicking, is what identifies the treatment effect of the groups. All participants were informed prior to clicking that they may be assigned to no group, so the control condition did not come as a surprise or involve any deception.[7]

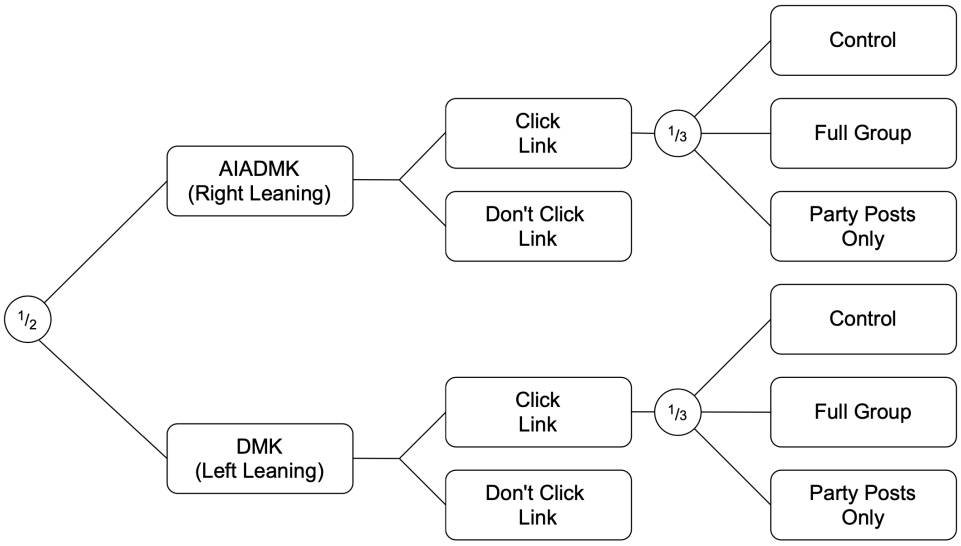
[5]Due to imperfect targeting of recruitment ads and nonresponse to the location question in the survey, some participants did not report voting in one of the 25 study constituencies. Each constituency is part of a larger voting district, and participants outside of a study constituency were randomly assigned to a study constituency within their district. Participants who reported voting in a district outside of the study area or who did not specify their district were randomly assigned to one of two groups organized at the state level.

[6]WhatsApp users typically come across an invitation link in isolation, often when shared on social media, and make a decision about joining that specific group.

[7]In principle, control participants could track down the sign-up links and sign up themselves. This would bias treatment effects toward zero. However, I observe the membership lists of all groups in the study and find that it is rare for control-group participants to join a treatment WhatsApp group.

Participants were assigned to a full group, a party-content-only group, or the control with equal probability. This randomization allows me to identify the causal effect of joining the different types of groups for the kinds of people who choose to join such a political WhatsApp group. Arguably, this is precisely the relevant population for such a causal inference exercise. Figure 1 provides a visual overview of the randomization process.

Figure 1: Experimental Design



Of all participants, 49%, clicked the link to join the randomly assigned group—a high join rate for an online experiment. As expected, participants were more likely to click the link to join a group if they preferred the party organizing the group. Figure 2 is a binned scatter plot with a linear regression fit, plotting join rates against a baseline political preference index. I define preference in terms of the randomly assigned party, with more positive numbers representing a greater relative baseline preference for that party. The index is normalized with a mean of zero and a standard deviation of one. The figure shows a positive, approximately linear relationship between relative preference for a party and the likelihood of clicking to join that party's group. Approximately 75% of participants who most like the assigned party click to join, while approximately 25% of participants who least like the assigned party click to join. The 49% of baseline participants who clicked their assigned link make up the experimental sample. Only the experimental sample received invitations

to the endline survey; those who did not click a link to join a group are "never-takers" who cannot contribute to the identification of treatment effects for the group they were assigned.

Because I randomized conditional on clicking a link to join a group, the average treatment effects that I estimate have the same local average treatment effect (LATE) interpretation as the treatment on the treated (TOT) effects that I would have estimated had I randomized without first conditioning on clicking.[8] The average treatment effects I estimate are averages for the groups of compliers who were willing to click a link. As Figure 2 shows, the compliers who click disproportionately prefer the party they are assigned at baseline. It is important to keep this in mind when interpreting the average treatment effects presented in the results that follow. To better understand these average effects, I also present heterogeneity throughout the analysis, separately estimating treatment effects by baseline preferences (the horizontal axis in Figure 2).

3.3 Data

Participants completed a baseline survey approximately three weeks before the election and an endline survey between the election and when the results were announced. Endline outcomes were registered in the American Economic Association's RCT Registry (AEARCTR-0007371).

Knowledge. To measure political knowledge, I administered a quiz at the endline. Each participant saw a series of news headlines and responded with their perceived likelihood that the headline was true using a five-point Likert scale. Some headlines were true, some were rumors circulating online and debunked by fact-checking websites, and some were false, written by research assistants in Chennai. Each participant saw one story of each type (true, rumor, and false) about each of the main state-level parties and the national ruling party,[9]

[8] It is also possible to back out the ITT that would have resulted by randomizing invitation links without first conditioning on compliance. Because never-takers would not have experienced any treatment effects, backing out an ITT is simply a matter of scaling my average treatment effects by the compliance rate of 0.49.

[9] The national ruling party, the Bharatiya Janata Party (BJP), only contested several seats in the 2021 Tamil Nadu Legislative Assembly election. However, they played a central role in the election, and whether

for a total of nine stories per participant. Participants knew that some stories would be true and some would be false but did not know the exact breakdown, and stories were chosen randomly from a list and presented in random order. To minimize the risk associated with exposing participants to untrue headlines, participants learned which stories were true and false immediately after completing the quiz. The primary outcome in this analysis is a score that sums a participant's confidence level in the true stories and subtracts their confidence level in the false stories and rumors.

Political Preferences. I measure political preferences at the baseline and endline using a series of Likert scale questions. I then group the questions and create z-score indices following Anderson (2008). This method first normalizes each input variable and then assigns weights using the inverse of their covariance matrix. The main index I use to measure political preferences takes four variables as inputs: seven-point Likert scale ratings of the two main parties and their leaders.[10] To measure preferences over policy issues, I created an index of questions about the two main parties' performance on salient policy issues, such as the high-profile farmer protests, education policy, employment and the economy, and the coronavirus response. I also elicited affective polarization—negative feelings toward members or supporters of the opposing party—using standard survey questions (Levy, 2021) and aggregated them into an index. All these questions appeared in the baseline and endline surveys. In the endline, I also measured self-reported voting.[11] In the analysis, I define these outcomes in terms of the randomly assigned party, with more positive values indicating stronger support for that party. All index variables have a mean of zero and a standard deviation of one, so treatment effects are measured in standard deviations.

Media Use. To detect "echo effects"—wherein social media exposure changes consumption of traditional media (Levy and Razin, 2019)—I collected a series of outcomes about media use.

or not certain parties and politicians favored the BJP was a widely discussed point of contention. The BJP is known for being active on social media.

[10]Both parties focused their campaigns on these leaders, who were also their candidates for Chief Minister, the state's executive.

[11]To verify voting, I asked participants to upload a photo of their left index finger, which is marked with indelible ink when they vote. Nonetheless, the party a participant voted for is entirely self-reported.

This included baseline and endline questions about time spent on and trust in traditional and social media. Participants also completed an endline module where they made choices about what news to consume and share. They saw three short video news clips headlines, truthfully attributed to a left-leaning, right-leaning, and centrist television station. The partisan clips each featured interviews with a politician from each respective party, projecting confidence about the election result. The centrist clip was a story about election turnout. Participants had a choice of which video to watch. The videos were embedded in the survey, and the duration viewed was recorded. After watching the video, participants had the option to share the video using a pre-filled WhatsApp message, and shares were recorded by redirecting the video through a unique link for each participant. This provides measures of observed behavior to supplement the self-reported survey questions. Outcomes from this module include which video was selected, the duration watched, and the number of observed shares.

WhatsApp Activity. WhatsApp allows all group members to view and export all posts to groups, including detailed information on precise times users join and leave groups. I match these data to study participants using the phone numbers they report. This allows me to observe the full extent of a participant's treatment, including whether they joined the assigned group, how long they stayed in the group, what kinds of posts they were exposed to, and how frequently they themselves posted. Observing the content of all posts allows me to explore mechanisms that distinguish individual and party messaging in Section 5. Additionally, WhatsApp's "seen by" feature permits measuring how quickly (if at all) a participant viewed a message posted to a group. I posted trivial messages to each group at three points throughout the study to observe how quickly a participant viewed the post.[12] This provides three precise audits of attention for each participant in the two treatment conditions.

[12]These audit messages were "good morning" or "good evening" images, which are common in these WhatsApp groups but have no material content.

3.4 Sample Recruitment and Timeline

Using Facebook ads, I recruited a sample of 3,056 adults living in Tamil Nadu in the month leading up to the election. The ads were geographically targeted and restricted to adults 18 years of age or older. The ads invited viewers to participate in a survey about current events. The ads did not mention politics, to avoid recruiting a sample selected on interest in politics. I used a variety of images in the ads, allowing Facebook to optimize image selection; the ad that most participants saw appears in Figure A1. Participants were told that if they completed the assigned surveys, they would have a chance to win a raffle for an iPhone, but there were no other direct benefits to participating. I screened out respondents who were not WhatsApp users or not adults, which were the only requirements to participate. I targeted the ads geographically at 25 constituencies around Chennai where I had links to party-organized WhatsApp groups at the constituency level for both leading parties.

Participants who clicked the ad were directed to a Qualtrics survey where they read a brief study description and consented to participate. Then, participants completed a self-administered baseline survey that took approximately 15 minutes. The randomization and intervention happened at the end of this survey. Of all participants, 96% completed the survey on a mobile phone, making the transition between Facebook, Qualtrics, and WhatsApp smooth and limited to a single device.[13] Of the 3,056 baseline participants, 1,491 (49%) clicked to join the WhatsApp group assigned to them and were randomized to one of the three treatment conditions. These 1,491 participants constitute the experimental sample.

The endline survey took place 26 days after the baseline, on average. Participants received an email invitation to participate in the endline on Qualtrics the day after the election. I followed up with participants by WhatsApp, email, and retargeted Facebook ads until the election results were announced four weeks later. Eventually, participants who did not complete a self-administered endline were called by a surveyor who attempted to

[13]WhatsApp is available as a desktop application and can be accessed through a web browser, so those who took the survey on a personal computer could also potentially participate seamlessly without switching to their phones.

administer the survey by phone. Of all endline surveys, 28% were conducted by phone, and the rest were self-administered. In total, 83% of participants in the experiment responded to the endline survey. For an online field experiment, this attrition rate is exceptionally low.

4 Experimental Results

The following analysis focuses on the 1,491 participants who reached the end of the baseline survey, accepted a group invitation link, and were assigned to a treatment condition. Table A1 shows that endline completion rates and baseline observables are balanced across treatment groups. It also characterizes the experimental sample. The sample is drawn from a wide distribution of ages, with a mean age of 41. Control-group participants are slightly younger than full-group participants. The sample is also highly educated. This is in part a feature of Tamil Nadu, which has a tertiary enrollment rate of over 50%, the highest among large Indian states (AISHE, 2020). Only 7% of the sample identify as female. The gender gap in social media use in developing countries is well documented and is especially large in South Asia (Fatehkia et al., 2018). Participants spend a large amount of time on WhatsApp, reporting over three hours of daily use on average. They also are more likely to say that they mistrust rather than trust information they see on WhatsApp. Other than age, there are no statistically significant differences between groups for any of these variables. Joint F-tests are not significant.

Throughout this section, I regress endline outcomes on treatment indicators for the full-group and party-content-only conditions. All regressions shown include a constant, district fixed effects, and baseline measures of the outcome variable where available. I present standard errors in parentheses and randomization-inference p-values in brackets (Young, 2019). All tables also include randomization inference p-values from a test of equality of the two treatment effect coefficients.

4.1 Knowledge

Much of the public discourse on the role of social media in politics focuses on its potential to foment inaccurate beliefs. Alternatively, a more optimistic hypothesis is that social media could increase political knowledge among a broader public as internet access increases.

To evaluate these alternatives, I measured participants' knowledge of political events at the endline. Each participant saw a series of news headlines and responded by assessing the likelihood that the headline was true using a five-point Likert scale. Three of the headlines were true, taken from stories in major newspapers during the study period. Three were rumors that circulated widely online and had been identified and debunked by fact-checking websites. Three were false and created by research assistants in Chennai. The inclusion of these false stories allows me to measure whether treated participants are more (or less) able to evaluate fake stories that they have no chance of having seen (in contrast to the effect of possible fake-news exposure in the groups). All three types of headlines primarily concerned parties' or politicians' statements about policy positions: this included central campaign promises reported from the parties' manifestos and statements political leaders made about widely discussed political issues. The primary outcome in this analysis is a score that sums the confidence level in all true stories and subtracts the confidence level in false stories and rumors.

Column 1 of Table 2 regresses this score on the treatment indicators. The full groups significantly improve members' ability to distinguish true from false news stories. Total scores increase by 0.82 points (or 0.18 standard deviations) on average due to the full groups. The point estimate on the party-content-only treatment is not statistically significant at conventional levels. The randomization inference p-value comparing the two treatment effects (p=0.085) allows me to reject their equality. Columns 2, 3, and 4 of Table 2 separately estimate treatment effects on confidence in true, rumored, and false news stories, respectively. In columns 3 and 4, I flip the sign of the outcome so positive coefficients correspond to an increase in knowledge (a decrease in confidence in rumors and false headlines). This shows

that most of the improvement in knowledge in column 1 comes from increased confidence in true headlines. Notably, belief in rumors also decreases, but the decrease is marginally significant. I further decompose these results by party in Table A2, separating stories by whether they pertain to the assigned party, the other party contesting the election, or the BJP. This reveals that most of the increase in ability to identify true news stories comes from stories about the assigned party. The point estimate of the treatment effects on beliefs in true stories about the other party is also positive but smaller and not significant at conventional levels. The groups also significantly reduce belief in rumors about the BJP. This is consistent with the campaigns' character and the groups' content, which highlighted BJP affiliation and positions as a major cleavage.[14]

4.2 Political Preferences

Table 3 shows results of regressions of political preference outcomes on indicators of the two treatment conditions. The outcomes in the first four columns are indices calculated using the inverse variance-covariance of their component variables as weights, following Anderson (2008). The variance-covariance weights are estimated from the full baseline sample, including participants who did not click the randomly assigned link. I use these weights to construct baseline and endline index values. All component variables are measured using five-point Likert scales. The outcome index in the first column uses questions about the two main parties and their leaders as inputs (the same index as in Figure 2). The index in the second column uses agree/disagree statements about divisive policy issues as inputs. The partisan affect index in column 3 comes from questions about affective polarization that are standard in the literature (Iyengar and Krupenkin, 2018). The outcome in column 4 is an indicator of whether the participant voted for the party whose group they joined. This was

[14]The BJP is widely unpopular in Tamil Nadu, as baseline survey responses show. In the election, the DMK tried to emphasize the AIADMK's national alliance with the BJP, which the AIADMK, in turn, tried to downplay. The BJP's policy stances were widely discussed in this election, despite their limited presence in the Assembly.

self-reported. To further verify voting status, I asked participants to upload a photo of their left index finger, which is marked with indelible ink when they vote.

The first coefficient in the first column of Table 3 shows that joining a WhatsApp group had a statistically significant impact on political preferences. On average, the full WhatsApp groups pushed their members' ideology 0.07 standard deviations in favor of the party affiliated with the group. In contrast, the parties' campaign messaging alone had no significant impact on persuasion. The difference between the treatment effects of the full groups and party-content-only groups is marginally insignificant at conventional levels. I find no statistically significant treatment effects on the other measures of political preferences and attitudes. Column 2 shows treatment effects on an index of policy issues that were highly discussed leading up to the election.[15] Point estimates of the effect of the full groups go in the expected direction, with treated individuals being pushed toward the policy stance of the assigned party. But the estimate is smaller than the estimate from column 1 and not statistically significant at standard levels. The estimated effect of the party-content-only treatment is smaller and also not significant. Column 3 shows treatment effects on political affect, using survey measures common in the literature.[16] Here, there are no statistically significant treatment effects of either intervention. This is notable, as previous experimental work on social media persuasion finds treatment effects on political affect but not opinions or beliefs (Levy, 2021).

Finally, column 4 shows treatment effects on stated voting behavior. The outcome is an indicator taking the value one if the participant reports having voted for the group randomly assigned at baseline and zero otherwise. Members of the full groups and party-content-only groups are 2.4 and 0.6 percentage points more likely to report voting for the randomly assigned party, respectively, but these estimates are not significant at conventional

[15]The list of issues includes education policy, stance on farmer protests, corruption, economic policy, COVID-19 response, cyclone response, and issues of Tamil identity and language.

[16]This includes five-point Likert scale agree/disagree statements, including "I think [party] followers have good ideas" and "I have a hard time seeing things from [party] members' point of view."

levels. Given the difficulty of influencing voting behavior and the small expected treatment effects, the study was not powered to detect treatment effects on voting.

Heterogeneity

Analyzing treatment effect heterogeneity is especially useful in this setting. In interpreting the average effects on political preferences, it is useful to know whether the groups influence baseline moderates or partisans more. Furthermore, there may be non-monotonic effects, in that group membership could generate backlash among participants who join the group of a party they oppose, pushing them further from that party. It is worth emphasizing that the average treatment effects discussed above are averages across my experimental sample, which by construction is the population of baseline participants willing to join a WhatsApp group for the party randomly assigned to them. As Figure 2 shows, this includes participants who strongly liked and strongly disliked the party assigned to them. To understand treatment effects across this distribution, I separately estimate treatment effects by baseline preferences.

To begin, I define five quintiles of the full baseline political preference distribution. I then sort participants into these quintiles based on their preferences for the party whose group they were randomly assigned. Participants in the first quintile are those whose preference for party assigned to them was at or below the twentieth percentile of the entire baseline distribution—those who like their assigned party the least. Participants in the fifth quintile are those whose baseline preference for party assigned to them was at or above the eightieth percentile of the entire baseline distribution—those who liked their assigned party the most. Defining the quintiles using the full baseline distribution (rather than the distribution of participants who joined a group) ensures that participants in the middle quintile are at the middle of the full distribution of preferences: moderates indifferent between the parties at the baseline. One drawback of this is that relatively few participants joined a group for a party they disliked, so there are fewer experimental participants in the lower quintiles, and the estimates for these quintiles are less precise.

Figure 3 separately plots treatment effects on political preferences for each quintile. Panel A plots treatment effects on the parties and leaders index by quintile of the baseline distribution. Panel B plots treatment effects on the political issues index, disaggregated by the same quintiles. Panels A and B both show that treatment effects on these outcomes are largely positive across the baseline distribution of preferences. Even participants in the lowest quintile, who disliked the party whose group they joined at the baseline, were pushed in the direction of that party. I cannot reject the hypothesis of null effects in these bottom quintiles but can rule out backlash effects of 0.2 standard deviations or more (in the direction of the opposing party) at the 95% level.[17] The second result from these panels is that persuasion is largely driven by participants at the middle of the baseline political preference distribution. These participants were relatively indifferent between the two parties at the baseline. Treatment effects for participants who more strongly favor the party assigned to them are smaller and less significant.[18] As with the average treatment effects, there are larger treatment effects among moderates assigned to the full, social groups than the party-content-only condition.

Panel C of Figure 3 shows larger point estimates on affective polarization at the bottom and middle of the baseline distribution (those who disagreed with the assigned group or were indifferent), but the estimates are imprecise. Panel D shows that the groups—particularly the full groups—move participants in the middle three quintiles toward voting for the assigned party, though only the fourth quintile's point estimate is significant at conventional levels. These point estimates—between 8 and 17 percentage points—are incredibly high for an outcome such as voting behavior, though the confidence intervals are wide.

4.3 Media Use

Next, I analyze treatment effects on media use. Column 1 of Table 4 shows that the full groups increase time spent on WhatsApp by an average of 16.5 minutes per day (SE 9.3).

[17] A standard deviation of 0.2 is approximately the magnitude of treatment effects on this outcome for baseline moderates.

[18] While few participants are at a true corner at the baseline (with the highest index value possible), that the index comes from Likert scale questions may mean that those at the extremes have less scope for moving along the index.

The point estimate on time use for the party-content-only condition is also positive but smaller and less significant, though I cannot reject the hypothesis that the coefficients on the two treatment indicators are equal. Columns 2 and 3 show that the treatment effects on self-reported trust in news on WhatsApp and the number of political messages forwarded are small and not significant at conventional levels.

Participants also completed an endline module where they made choices about what news to consume and share. They were presented with three short video news clip headlines, attributed to a left-leaning, right-leaning, and centrist television station. Column 4 shows that treated participants were no more likely than the control to choose the news station corresponding to the group assigned to them. Column 5 shows that the duration of news watched (measured in seconds) was not significantly different between the full-group condition and the control. Members of the party-content-only groups watched 11.6 seconds more of partisan news relative to the control. Column 6 shows that the treatment also had a null effect on the likelihood of sharing the viewed link on WhatsApp. The number of observations in columns 4 through 6 is smaller because participants who took the endline survey by phone call did not complete this module, which involved viewing videos embedded in the survey platform. These results do not suggest any "echo effects" of the groups on their members' self-reported or observed media use.

5 Mechanisms

These experimental results suggest that the horizontal messaging between group members is crucial to the groups' effects, as party messaging alone has consistently smaller and less significant effects than the full groups. But what drives this difference? To understand how the two treatments differ, I use administrative WhatsApp data. Full chat histories exported from the study groups allow me to use all group activity as data. I combine this with data from audit messages I posted to the groups, allowing me to measure participants' attention

using WhatsApp's "seen by" feature. I use these data to understand the mechanisms behind the differential effects of the two treatment conditions.

Several possible mechanisms that distinguish the full-group condition from the party-content-only condition. First, the full groups receive a substantially higher volume of messages (sent by individuals, as both groups have identical party messages). The relationship between message volume and treatment effects is theoretically ambiguous. More group activity could result in greater engagement, and the mere repetition of information could increase impacts. But message volume could also crowd out attention. People inundated with messages in a WhatsApp group may be more likely to stop checking, mute, or even leave the group. The relationship between message volume and treatment effects may not be monotonic: more messages may initially increase treatment effects but decrease them after a certain point by crowding out attention. The administrative data from WhatsApp allow me to directly observe both message volume and attention.

Other mechanisms might also cause the full groups to have larger effects than the party-content-only groups. The content and sentiment of messages in the two kinds of groups may differ in important ways. Being able to post oneself may be an important feature of the full WhatsApp groups since active participation might drive the effects. People may also make different inferences about a message depending on its source. Messages from individuals may also influence perceptions of norms, which can, in turn, impact preferences. There may be interactions between these mechanisms, which are not mutually exclusive. The experiment was not designed to generate exogenous variation in all these mechanisms. In this section, I focus on the two mechanisms where my data can provide clear insights into how the treatment conditions differ: volume and attention. I also briefly discuss alternative mechanisms and how they might be identified in future experimental work.

5.1 Volume

To study the role of message volume, I analyze treatment effect heterogeneity by WhatsApp groups. I rely on the natural variation in the number of individual and party messages sent to

the WhatsApp groups during the study period. I find that full groups receiving more party or total messages do not have larger treatment effects. However, there is suggestive evidence that in the party-content-only condition, participants who received more messages (from the party) were more strongly persuaded. This is consistent with decreasing returns to message volume, where additional messages increase persuasion at the low end of the support (as in the party-content-only condition) but eventually taper off once participants receive a large volume of messages (as in the full-group condition).

First, I focus on the volume of messages coming specifically from party administrators. I create an indicator that identifies groups that had an above-median number of messages from party administrators during the study period.[19] I add this indicator to the main regressions and interact it with the two treatment indicators. The coefficients on these interaction terms represent the additional treatment effect in groups with an above-median number of party messages (compared to groups with a below-median number of party messages).

The first coefficient in panels A and B of Figure 4 shows the relationship between party-message volume and treatment effects in the full groups. These coefficients are not significantly different from zero, meaning that full groups receiving an above-median number of party messages do not have significantly larger treatment effects on knowledge or preferences. The second coefficient in panels A and B shows this same interaction, but for party-content-only groups. Panel A shows that there is no significant relationship between message volume and knowledge treatment effects in the party-content-only groups. However, the positive coefficient in panel B shows that party-content-only groups receiving more messages have higher average treatment effects on political preferences. This difference is marginally insignificant (p=0.11).

Next, I focus on the total volume of messages sent to the WhatsApp groups, summing the number of messages sent to each group from individuals and party administrators over the

[19] In the party-content-only groups, this is also the total number of messages to the group, as no messages were shared from individuals. I also assign the indicator to the control groups; in which case, it represents the volume of party messages in the group the participant would have been invited to.

study period. Because party-content-only groups did not receive messages from individuals, I only show treatment effect heterogeneity by total message volume for the full groups. Again, I create an indicator for being assigned a group with above-median volume, this time using the total volume of messages. The rightmost coefficients in Figure 4 correspond to the interaction between this total volume indicator and the full-group treatment variable. These coefficients are not significantly different from zero, meaning that full groups with above-median total volume do not have greater treatment effects.

Taken together, these results show that greater message volume does not typically correspond to greater treatment effects. One exception to this may be the less active party-content-only groups, where groups with more messages have greater treatment effects on political preferences. This is consistent with diminishing returns to message volume, where additional messaging is persuasive only when the initial messaging volume is low.

5.2 Attention

To compare attention to full and party-content-only groups, I rely on a series of attention audits. WhatsApp's "seen by" feature allows anyone who posts to a group to check which group members visited the group and saw their message. To exploit this feature, I posted to the groups at three points during the study. A goal of these audit messages was to engage in a naturalistic way, not disrupting the group's discourse, while also not contributing any political content. Most of group members' posts related to politics and the election, with one notable exception: group members often posted images saying "good morning" or "good evening" to each other. These messages are common in WhatsApp groups across India.[20] This allowed me to post to each group at three points during the study and observe participants' attention to my messages without changing the nature of the intervention in a meaningful way. I recorded the phone numbers that had seen the messages within 48 hours of their posting and matched them to the phone numbers participants provided in the baseline.

[20]These messages have been so common that the Wall Street Journal reports they have posed major data storage constraints for many Indian smartphone users. Prime Minister Modi reportedly sends "good morning" messages and has admonished lawmakers for not responding to these messages. (Purnell, 2018)

Figure 5 uses these audits to compare attention in the full and party-content-only groups. The means plotted are the mean fraction of audits that participants in each treatment condition viewed within 48 hours. This shows that for a given message, participants in the full groups paid more attention than in the party-content-only groups. Importantly, this is unconditional on having stayed in the group to actually receive the audit message. Full-group members were more likely to view a message posted to their assigned group, despite receiving more messages and leaving their groups earlier, on average. This suggests that even though the volume of messages might be off-putting to some full-group members, causing them to leave, those who stayed in the full groups ended up checking it more as a result of the individual messages, such that overall attention in the full groups is higher than in the party-content-only groups.

5.3 Additional Mechanisms

In addition to volume and attention, other mechanisms may distinguish the full-group and party-content-only conditions. The posts of parties and individuals may have different content. For example, peers sometimes fact-check each other's posts, which experimental research has shown can reduce belief in misinformation (Badrinathan et al., 2020). Participants' ability to post themselves to the full groups may be an important factor that increases engagement. One in four participants assigned to a full group posted to the group at some point during the study, demonstrating the attractiveness of this feature. People may also make different inferences about a message depending on its source. On the one hand, parties may be more credible sources of information than unknown individuals. On the other hand, messages from individuals may be perceived as more authentic or less obviously partisan than messages from a political party. Messages from individuals may also influence perceptions of norms and the beliefs of others (second-order beliefs). Bursztyn et al. (2020) show that perceived norms can impact stated political preferences and choices. These mechanisms are not mutually exclusive and may interact. In the context of this experiment, it would not have been practical or naturalistic to exogenously vary these mechanisms individually.

However, it would be possible in future lab or field experiments to vary some of these features of social media communication in isolation, such as knowledge about a message's source or the ability to respond, to measure how they influence belief and preference updating. Better understanding the determinants of people's responsiveness to social media messaging is a policy-relevant topic that merits further research.

6 Conclusion

Social media is increasingly a source of political information for voters in developing countries. In this paper, I show that political groups on the messaging platform WhatsApp increase voter knowledge, improving users' ability to distinguish true from false news. This is striking, given the public concern about political misinformation, especially on WhatsApp in India. One caveat to consider when interpreting these study results is the context. Party-organized WhatsApp groups are common around the world and an important object of study. However, there are other ways that political information is shared on WhatsApp, such as political groups organized by individuals, nonpolitical groups, and individual chats. Misinformation may be more pervasive in these contexts than in groups created by party officials. Nonetheless, the present study demonstrates that focusing only on the drawbacks of political engagement on social media may be misguided. Understanding whether the findings of this study generalize to settings with lower information quality is an important topic of future research.

This paper also shows that groups organized by political groups on WhatsApp can influence people's stated political preferences. Notably, this result comes primarily from baseline moderates being pushed toward the party whose group they join. Rather than galvanizing the base of supporters, the content in the group persuaded moderates. Considering concerns about social media fomenting extremism, this result is also encouraging. Although group membership pushed moderates away from the middle, participants at either extreme of the baseline preference distribution were largely unmoved. Importantly, there was no significant

increase in affective polarization. This is in contrast to evidence on the effects of Facebook in the United States, where affective polarization is the primary outcome impacted by political information (Levy, 2021).

This paper provides new evidence on a class of platforms that are understudied in the literature on social media and politics: messaging applications. These applications, particularly WhatsApp, are especially popular in developing countries. WhatsApp has several fundamental features that distinguish it from platforms like Facebook and Twitter. First, no algorithm governs the order in which users see posts. If people are more likely to engage with surprising or inflammatory posts, an algorithm could be more likely to feature such content in user feeds. On WhatsApp, this is of less concern. WhatsApp communication may also operate on a smaller scale. Group size is limited to 256 people, such that the number of people WhatsApp users can interact with is typically much smaller than Facebook or Twitter. Furthermore, WhatsApp messages are end-to-end encrypted, which may influence how users communicate on the platform.

Understanding whether these features play a pivotal role in the platforms' respective treatment effects is a topic for future research—and it has implications for the debate on social media regulation. For example, removing Facebook's algorithm and replacing it with a reverse-chronological feed—like on WhatsApp—is a proposal that has been discussed in response to recent whistleblower claims (Zorthian, 2021). Facebook recently introduced an option where users can choose to turn off algorithmic ranking in favor of a reverse-chronological feed (Sethuraman, 2021). In India, the Central Government has asked WhatsApp to break its end-to-end encryption, allowing government officials to view and regulate content, which WhatsApp has challenged in court (Menn, 2021). Changing any of these features is likely to impact how people communicate and the content they are exposed to. However, it is difficult to predict the result of such changes without understanding what motivates people to communicate on social media and how they draw inferences from the communication they see. Further research on these topics would help inform policy debates.

References

AISHE (2020). "All India survey on higher education". Technical report, Indian Ministry of Education.

Aker, J. C., P. Collier, and P. C. Vicente (2017). "Is information power? Using mobile phones and free newspapers during an election in Mozambique". *Review of Economics and Statistics 99*(2), 185–200.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020). "The welfare effects of social media". *American Economic Review 110*(3), 629–76.

Anderson, M. L. (2008). "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects". *Journal of the American Statistical Association 103*(484), 1481–1495.

Badrinathan, S., S. Chauchard, and D. Flynn (2020). "'I don't think that's true, bro!': An experiment on fact-checking misinformation in India". Technical report, Working Paper.

Banerjee, A., M. Alsan, E. Breza, A. G. Chandrasekhar, A. Chowdhury, E. Duflo, P. Goldsmith-Pinkham, and B. A. Olken (2020). "Messages on COVID-19 prevention in India increased symptoms reporting and adherence to preventive behaviors among 25 million recipients with similar effects on non-recipient members of their communities". Technical report, National Bureau of Economic Research.

Banerjee, A., S. Kumar, R. Pande, and F. Su (2011). "Do informed voters make better choices? Experimental evidence from urban India". *Unpublished manuscript*.

Bharadwaj, D. (2021). "Blocked 2 million accounts in India in 1 month, says WhatsApp in new report". *Hindustan Times*.

Bidwell, K., K. Casey, and R. Glennerster (2020). "Debates: Voting and expenditure responses to political communication". *Journal of Political Economy 128*(8), 2880–2924.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012). "A 61-million-person experiment in social influence and political mobilization". *Nature 489*(7415), 295–298.

Bursztyn, L., G. Egorov, and S. Fiorin (2020). "From extreme to mainstream: The erosion of social norms". *American Economic Review 110*(11), 3522–48.

Bursztyn, L., I. Haaland, A. Rao, and C. Roth (2021). "Disguising prejudice: Popular rationales as excuses for intolerant expression". Technical report, Working Paper.

Da Silveira, B. S. and J. M. De Mello (2011). "Campaign advertising and election outcomes: Quasi-natural experiment evidence from gubernatorial elections in Brazil". *The Review of Economic Studies 78*(2), 590–612.

Fatehkia, M., R. Kashyap, and I. Weber (2018). "Using Facebook ad data to track the global digital gender gap". *World Development 107*, 189–209.

Ferraz, C. and F. Finan (2008). "Exposing corrupt politicians: The effects of Brazil's publicly released audits on electoral outcomes". *The Quarterly Journal of Economics 123*(2), 703–745.

Fujiwara, T. and L. Wantchekon (2013). "Can informed public deliberation overcome clientelism? Experimental evidence from Benin". *American Economic Journal: Applied Economics 5*(4), 241–55.

Garimella, K. and D. Eckles (2020). "Images and misinformation in political groups: Evidence from WhatsApp in India". *arXiv preprint arXiv:2005.09784*.

George, S., S. Gupta, and Y. Neggers (2020). "Coordinating voters against criminal politicians: Experimental evidence from India".

Guriev, S., N. Melnikov, and E. Zhuravskaya (2021). "3g internet and confidence in government". *The Quarterly Journal of Economics 136*(4), 2533–2613.

Iyengar, S. and M. Krupenkin (2018). "The strengthening of partisan affect". *Political Psychology 39*, 201–218.

Jensen, R. (2007). "The digital provide: Information (technology), market performance, and welfare in the South Indian fisheries sector". *The Quarterly Journal of Economics 122*(3), 879–924.

Jiménez-Durán, R. (2021). "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter". Technical report, Working Paper.

Kumar, S. and P. Kumar (2018). "How widespread is WhatsApp's usage in India?". *LiveMint*.

Levy, G. and R. Razin (2019). "Echo chambers and their effects on economic and political outcomes". *Annual Review of Economics 11*, 303–328.

Levy, R. (2021). "Social media, news consumption, and polarization: Evidence from a field experiment". *American Economic Review 111*(3), 831–70.

Marx, B., V. Pons, and T. Suri (2021). "Voter mobilisation and trust in electoral institutions: Evidence from Kenya". *The Economic Journal 131*(638), 2585–2612.

Menn, J. (2021, May). "WhatsApp sues Indian government over new privacy rules - sources".

Murgia, M., S. Findlay, and A. Schipani (2019). "India: The WhatsApp Election". *Financial Times 5*.

Pande, R. (2011). "Can informed voters enforce better governance? Experiments in low-income democracies". *Annu. Rev. Econ. 3*(1), 215–237.

Perrigo, B. (2019). "How volunteers for India's ruling party are using WhatsApp to fuel fake news ahead of elections". *Time Magazine*.

Purnell, N. (2018). "The Internet is filling up because Indians are sending millions of 'good morning' texts". *Wall Street Journal*.

Ravallion, M., D. van de Walle, P. Dutta, and R. Murgai (2015). "Empowering poor people through public information? Lessons from a movie in rural India". *Journal of Public Economics 132*, 13–22.

Renno, R. (2019). "WhatsApp: The widespread use of WhatsApp in political campaigning in the Global South". *Our Data Our Selves*.

Sethuraman, R. (2021, Apr). "More control and context in news feed".

Young, A. (2019). "Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results". *The Quarterly Journal of Economics 134* (2), 557–598.

Zhuravskaya, E., M. Petrova, and R. Enikolopov (2020). "Political effects of the internet and social media". *Annual Review of Economics 12*, 415–438.

Zorthian, J. (2021). "Washington wants to regulate Facebook's algorithm. That might be unconstitutional". *Time Magazine*.

Tables and Figures

Table 1: Treatment Description

	Full group mean	Party posts mean
Daily posts received	113.6	7.4
Image	67.7	3.7
Text	26.0	1.1
Video	11.6	2.4
Link	6.7	0.1
Document (GIF, audio, etc.)	1.58	0.06
Group size	181	—
Ever posted	0.25	—
Number of posts	6.44	—
Observations	495	475

Notes: Table 1 summarizes participants' experience in the treatments. Summary statistics come from exported chat histories for the groups in the study. Means are calculated across study participants (not groups).

Table 2: Treatment Effects: Knowledge

	Total Score	By Story Type		
	(1)	(2)	(3)	(4)
		True	Rumor	False
Full group	0.816	0.490	0.259	0.063
	(0.325)**	(0.218)**	(0.188)	(0.177)
	[0.003]***	[0.007]***	[0.114]	[0.659]
Party content	0.331	0.327	-0.163	0.163
only	(0.325)	(0.225)	(0.194)	(0.170)
	[0.254]	[0.090]*	[0.337]	[0.288]
$\beta_1 = \beta_2$ p-value	0.085*	0.392	0.010**	0.501
Control mean	1.91	0.69	0.75	0.47
Observations	1,119	1,119	1,120	1,120

Notes: Table 2 presents treatment effects on knowledge, measured by a news headline quiz. The independent variables displayed are indicators for the two treatment conditions. The outcome in column 1 is the raw score on this quiz. Columns 2-4 decompose this treatment effect by type of message, with positive coefficients representing greater knowledge. All regressions include a constant and district fixed effects. Standard errors are in parentheses, and randomization-inference p-values are in brackets. The $\beta_1 = \beta_2$ p-values shown are randomization inference p-values from a test of the hypothesis that the effect of the two treatment conditions is equal.

Table 3: Treatment Effects: Political Preferences

	Inverse Covariance Indices			
	(1) Political Preference	(2) Policy Issues	(3) Partisan Affect	(4) Voted for Assigned Party
β_1: Full group	0.072	0.052	-0.007	0.024
	(0.039)*	(0.046)	(0.059)	(0.036)
	[0.029]**	[0.206]	[0.889]	[0.440]
β_2: Party	0.021	0.001	0.050	0.006
content only	(0.039)	(0.048)	(0.058)	(0.036)
	[0.535]	[0.988]	[0.317]	[0.845]
$\beta_1 = \beta_2$ p-value	0.136	0.198	0.251	0.582
Control mean	0.34	0.31	0.24	0.52
Observations	1,150	1,154	1,137	1,135

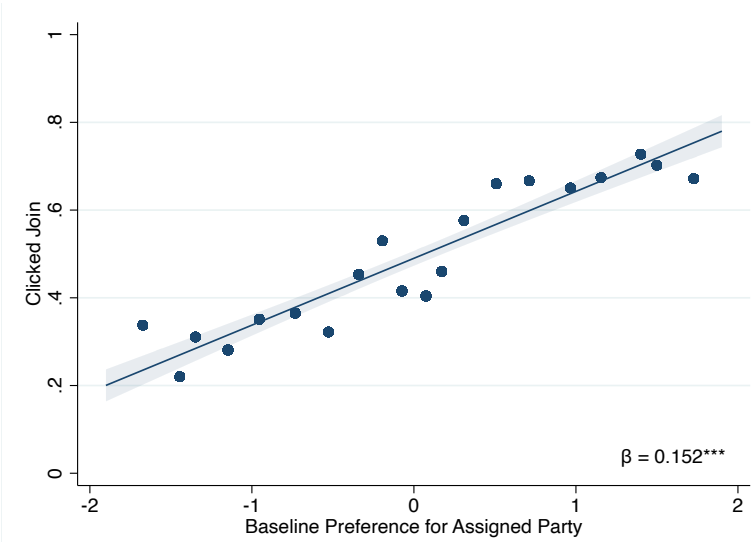
Notes: Table 3 presents treatment effects on indices of political preferences. The independent variables displayed are indicators for the two treatment conditions. All outcome indices are constructed following Anderson (2008) from Likert scale variables. The index in column 1 takes ratings of the two main parties and their leaders as inputs. The index in column 2 combines a series of agree/disagree questions about policy issues emphasized in the election. The index in column 3 takes common measures of affective polarization as inputs. The output in column 4 is an indicator of self-reported voting for the assigned party. All regressions include a constant, the baseline value of the outcome (except for voting), and district fixed effects. Standard errors are in parentheses, and randomization-inference p-values are in brackets. The $\beta_1 = \beta_2$ p-values shown are randomization inference p-values from a test of the hypothesis that the effect of the two treatment conditions is equal.

Table 4: Treatment Effects: Social Media Use

	Self-Reported			Observed Choices		
	(1) Daily Mins on WA	(2) Trust News on WA	(3) Fwds Last 2 Weeks	(4) Chose Partisan News	(5) Duration News Watched	(6) Shared News Link
β_1: Full group	16.5	0.003	2.66	-0.002	-0.124	-0.027
	(9.28)*	(0.084)	(2.46)	(0.035)	(1.528)	(0.038)
	[0.050]*	[0.960]	[0.254]	[0.944]	[0.999]	[0.410]
β_2: Party	8.69	0.055	2.50	-0.025	11.6	-0.026
content only	(9.59)	(0.085)	(2.49)	(0.035)	(9.95)	(0.038)
	[0.293]	[0.461]	[0.260]	[0.436]	[0.033]**	[0.416]
$\beta_1 = \beta_2$ p-value	0.311	0.469	0.934	0.437	0.025**	0.982
Control mean	177	-0.49	22.2	0.23	6.64	0.31
Observations	1,151	1,121	1,156	848	878	866

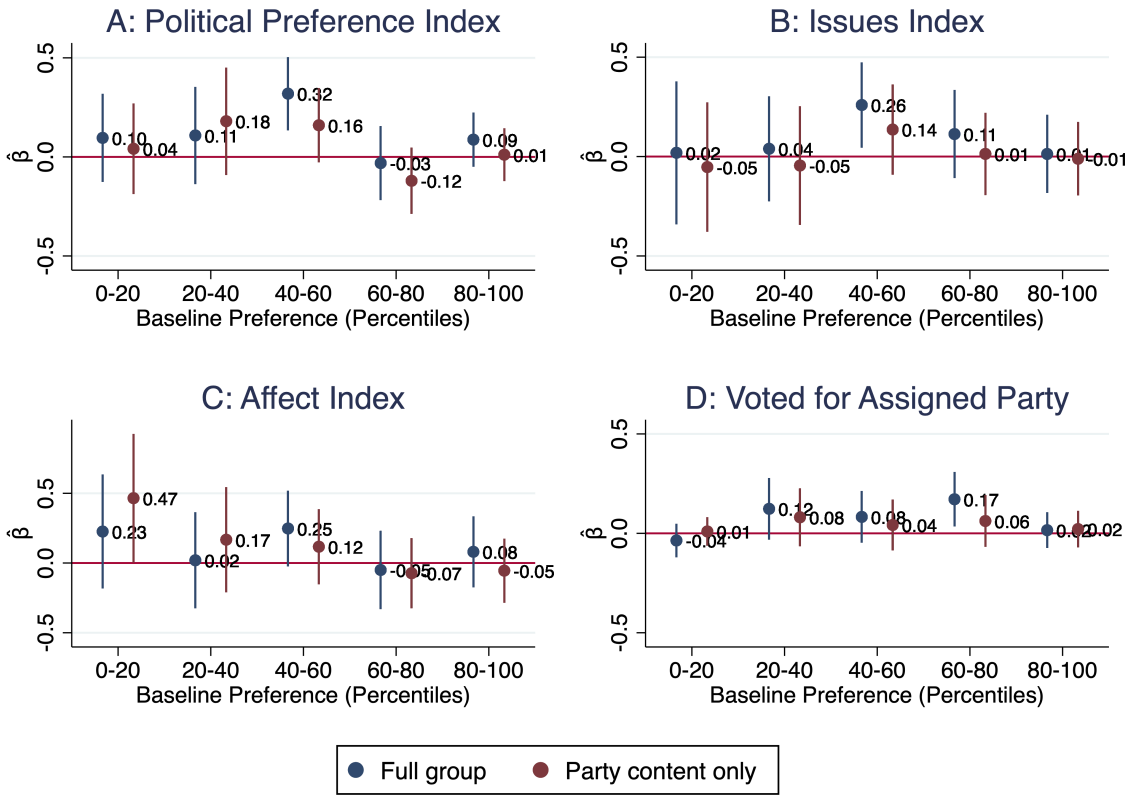
Notes: Table 4 presents treatment effects on social media use. The independent variables displayed are indicators for the two treatment conditions. Columns 1-3 have self-reported outcomes, and columns 4-6 have outcomes that come from observed choice in a module where participants made a choice of news to watch and share on social media. All regressions include a constant, the baseline value of the outcome (except for the revealed preference module, which was only conducted at the endline), and district fixed effects. Standard errors are in parentheses, and randomization-inference p-values are in brackets. The $\beta_1 = \beta_2$ p-values shown are randomization inference p-values from a test of the hypothesis that the effect of the two treatment conditions is equal.

Figure 2: Group Take-Up by Baseline Political Preference



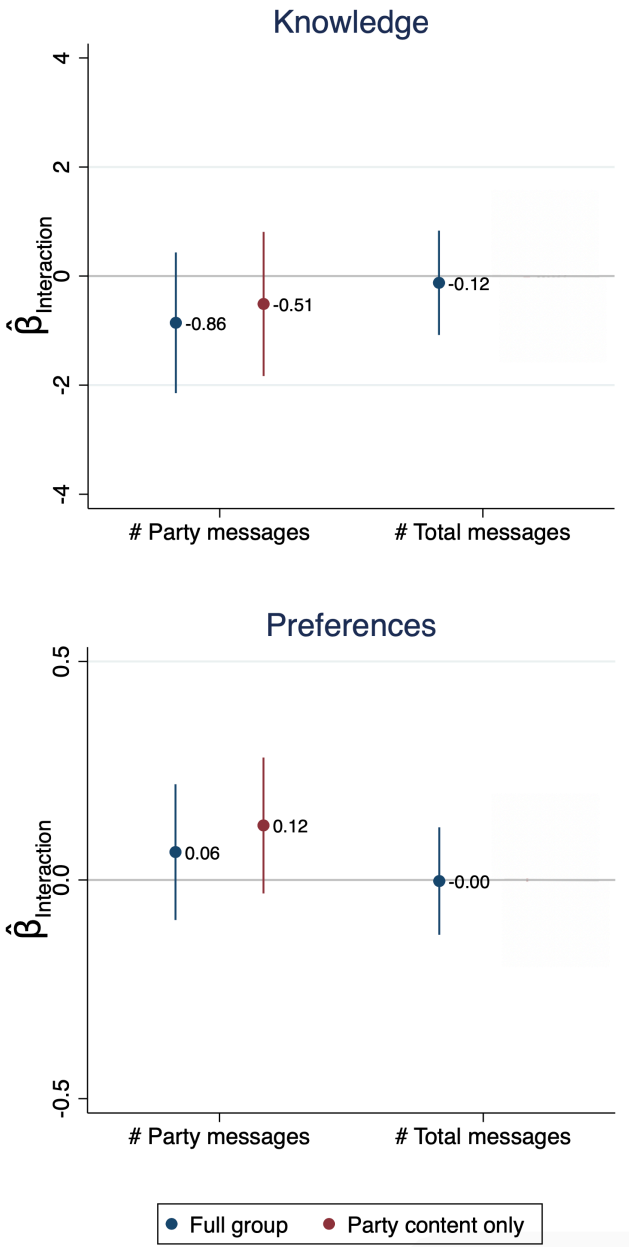
Note: Figure 2 plots WhatsApp group join rates against baseline political preferences. Baseline preference is defined in terms of the affiliation of the WhatsApp group offered, with more positive numbers indicating a stronger support of that party relative to the opposing party. This variable is an index of support for the two main parties and their leaders. The figure shows a binned scatter plot as well as a linear regression fit. Participants who like the party assigned to them are more likely to accept an invitation to its WhatsApp group: a one standard deviation increase in relative preference of the assigned party increases the likelihood of joining their group by 15.2 percentage points, on average. Those who clicked join were randomized to a group or the control and constitute the experimental sample. Thus, this figure characterizes the baseline preferences of the group of experimental participants represented in the average treatment effects I estimate.

Figure 3: Treatment Effect Heterogeneity: Baseline Ideology



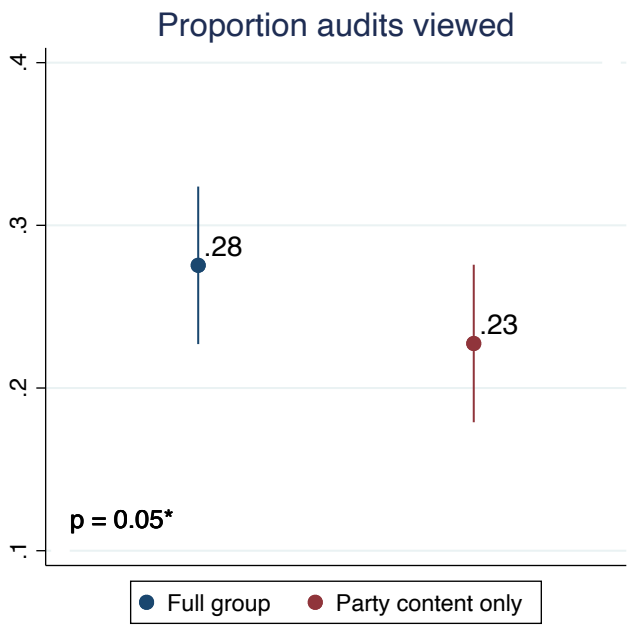
Note: Figure 3 shows treatment effect heterogeneity for the political preference outcomes in Table 3, plotted by baseline political preference. I separately estimate treatment effects by quintile of the baseline political preference distribution, defining quintiles in terms of the assigned party, with greater values indicating a stronger baseline preference for that party. Bars represent 95% confidence intervals.

Figure 4: Treatment Effect Heterogeneity: Message Volume



Note: Figure 4 shows treatment effect heterogeneity by group volume. Panel 1 regressions use the knowledge outcome from column 1 of Table 2 as the dependent variable, and panel 2 regressions use the preference outcome from column 1 of Table 3 as the dependent variable. The coefficients plotted correspond to the interactions term between the treatment indicators and an indicator for being in a group with above-median number of messages. This shows how much greater treatment effects are in groups with above-median message volume. I define the message volume above-median cutoff three ways: in terms of the total number of messages, the number of individual messages, and the number of party messages.

Figure 5: Attention by Group Type



Note: Figure 5 compares attention in the full groups and party-content-only groups. Attention is measured using a series of audit messages, posted to the groups three times during the study, and WhatsApp's "seen by" feature. The outcome is the fraction of audit messages viewed within 48 hours. The p-value shown is from a test of the difference between the groups, including controls.

A Appendix

Table A1: Treatment Balance

	(1)		(2)		(3)		T-test		
	Control		Full Group		Party content only		P-value		
Variable	N	Mean/SE	N	Mean/SE	N	Mean/SE	(1)-(2)	(1)-(3)	(2)-(3)
Endline completion	521	.848	495	.812	475	.836	.124	.587	.334
		(.0157)		(.0176)		(.017)			
Age	514	40.3	494	41.8	471	41.1	.0672*	.375	.343
		(.571)		(.602)		(.582)			
Education	515	12.7	492	12.7	467	12.7	.855	.846	.699
		(.044)		(.0443)		(.0393)			
Female	418	.0742	411	.0657	379	.0686	.633	.761	.871
		(.0128)		(.0122)		(.013)			
Took Survey in Tamil	521	.931	495	.921	475	.922	.555	.595	.959
		(.0111)		(.0121)		(.0123)			
Daily WhatsApp Hours	520	3.33	494	3.19	473	3.28	.41	.798	.59
		(.117)		(.114)		(.125)			
Trust Info on WhatsApp	510	-.241	486	-.208	463	-.246	.677	.95	.637
		(.0564)		(.0567)		(.0583)			
Political Preference Index	513	.341	487	.276	471	.304	.28	.545	.644
		(.0431)		(.0429)		(.0441)			
Policy Issues Index	520	.248	494	.284	474	.255	.567	.914	.647
		(.0437)		(.0456)		(.0451)			
Partisan Affect Index	513	.274	491	.265	469	.259	.878	.812	.933
		(.0452)		(.0446)		(.0434)			
Political Involvement Index	511	1.59	483	1.58	463	1.57	.916	.679	.755
		(.0357)		(.0353)		(.0367)			
F-test of joint significance (F-stat)							1.53	.351	.641
F-test, number of observations							787	750	747

Note: The value displayed for t-tests are p-values. The value displayed for F-tests are the F-statistics. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table A2: Treatment Effects: Knowledge, by Party

	Assigned Party			Other Party			BJP		
	(1) True	(2) Rumor	(3) False	(4) True	(5) Rumor	(6) False	(7) True	(8) Rumor	(9) False
β_1: Full group	0.30***	0.05	-0.01	0.15	-0.03	0.05	0.06	0.28**	0.02
	(0.11)	(0.10)	(0.10)	(0.12)	(0.10)	(0.10)	(0.11)	(0.11)	(0.09)
β_2: Party	0.17	-0.10	0.01	0.19	-0.11	0.06	0.05	0.13	0.09
content only	(0.11)	(0.10)	(0.10)	(0.12)	(0.10)	(0.10)	(0.11)	(0.11)	(0.09)

Note: Table A2 further decomposes the treatment effects on voter knowledge by both the type of message (true, rumor, or false) and the party the message pertains to. This reveals that the increase in knowledge is driven by an increased belief in true stories about the assigned party and a decrease in belief in rumors about the BJP, the national ruling party.

Figure A1: Recruitment Ad

Note: Figure A1 shows the ad that most participants saw when they were recruited to the study.